Department of Forest Genetics and Plant Physiology Swedish University of Agriculture Sciences, Umeå, Sweden Master degree thesis in Plant and forest biotechnology 30ECTS

A Next Generation sequencing approach to constructing a genetic map for *Populus tremula*

Tahir Mujtaba

Supervisor: Nathaniel Street Umeå Plant Science Center Department of Plant Physiology Umeå University SE-901 87 Sweden

2013

Department of Forest Genetics and Plant Physiology
Swedish University of Agriculture Sciences Umeå, Sweden
Master degree thesis in Plant and forest biotechnology 30ECTS

A Next Generation sequencing approach to constructing a genetic map for *Populus tremula*

Tahir Mujtaba

Supervisor: Nathaniel Street Umeå Plant Science Center
Department of Plant Physiology
Umeå University SE-90187 Sweden

Examinator: Ewa Mellerowicz
Swedish University of Agricultural Sciences
Department of Forest Genetic and Plant Physiology
Umeå

Plant and forest biotechnology
Master thesis
Cource code: EX0634,
Master thesis in Biology at the Department of Forest Genetics and Plant Physiology
Masterprogram 120 p
2013

Keywords: genomic polymorphism, *Populus tremula*

Contents

Acknowledgment

I would like to express my gratitude to my supervisor **Nathaniel Street** for proposing and supervising me for this project, his support, timely guidance and helpful criticism concerning the application of bioinformatics tools, results, presentations and the manuscript. I am very much thankful to him for his informative, enjoyable and fruitful discussions, also for Friday afternoons NGS group meetings providing a nice platform for helpful discussions and sharing different ideas and to SLU, Umeå University, UPSC and the fantastic Master program who made me want to stay.

To Nicolas Delhomme, for additional support and help in understanding and applying the Bioinformatics tools, for supporting me with setting up my project plan, discussing strategies and results and for thoroughly reading and commenting the manuscript. Also, I would like to thank him for his patience, for enlightening explanations, interesting and amusing discussions, which encouraged me to complete this project.

To Chanaka for introducing me to the Galaxy (local server/tool kit) and helping me to get used to the Linux operating system as well as solving the entire computer based problems for this project work. To Ioana Gaboreanu for all the support, I needed for wet lab analyses. To Jeanette Tångrot who spared her precious time for informative discussions and timely guidance.

To Ahmed Aley for introducing me to the world of Python, Perl and letting me understand entire computational world.

To my grandparents and my parents who passed their passion to me. Thank you all for always being there for me and Swedish University of Agriculture Sciences, UPSC for offering such a fantastic degree program and to all those who made a memorable stay at Umeå.

Tahir Mujtaba

Summary

Plant biologists have long been studying phenotypic and physiological variation and the molecular mechanisms underlying natural variation in these processes have stimulated scientists to uncover the genomic polymorphism responsible for such variation. *Populus tremula*, (European aspen, hereafter referred to as aspen), is a member of *Populus* genus, which has become a model system for genetic and genomic studies and, more recently, for studies linking genomics to ecology and evolution. Many *Populus* species can be efficiently genetically transformed, all have relatively small genomes of ~500 Mbp and a number of genetic maps have been constructed using various F_1 and more advanced crossing designs. Importantly, there is also a reference genome sequence available for *P.trichocarpa* (Tuskan *et al.*, 2006), which has significantly advanced the utility of *Populus* as a model system (Wullschleger *et al.*, 2012).

Next generation sequencing (NGS) techniques have made genomic studies significantly faster and easier to conduct due to their massively parallel and rapid generation of high-quality sequence data at relatively low cost per base pair. As a result NGS has rapidly become the technology of choice for most scientist conducting sequence projects. NGS has also revolutionized the field of gene expression analysis by its application to sequencing cDNA to assay gene expression levels and is increasingly replacing the use of expression microarrays in genomics studies.

This study compared the suitability of two methods for identification of polymorphism that could be used for constructing a genetic map to complement the aspen genome project by facilitating orientation and location of assembly scaffolds within the reference *P. trichocarpa* genome sequence (Tuskan *et al.*, 2006). We first assessed the applicability of utilizing a reduced representation sequencing library approach in the parents of an F_1 intraspecific *P. tremula* population (RRL). RRL construction involves the use of restriction enzyme digestion of genomic DNA to generate a consistent set of DNA fragment from different samples and serves the purpose of efficiently reducing genomic representation and subsequent volume of sequence data that is required for polymorphism identification.

The second focus of this study was to assess the use of existing gene expression data that had been generated using RNA-Sequencing (RNA-Seq) to generate a *de novo* reference transcript assembly for one of the parents of the *P. tremula* F_1 population. This reference assembly was then used for Single Nucleotide Polymorphism (SNP) detection using RNA-Seq data from

both parents to identify a set of SNPs would then be suitable for genotyping the F1 progeny to allow construction of a genetic map.

1. Introduction

About 30 species of poplar and aspen are colonized throughout the Northern Hemisphere, and there are substantial areas of planted Euroamerican poplars and inter-American hybrids of aspen in Europe, Asia and North America (Rinaldi *et al.,* 2007 and Paolucci *et al.,* 2010). *Populus* was adopted as a model system for forest geneticists and biologists due to the economical and ecological importance of many poplar and aspen species. Worldwide, *Populus* hold significant economic values and from a scientific perspective there are protocols for vegetative propagation, efficient genetic transformation, propagated material displays very rapid growth and there are a number of genetic maps available, all of which have contributed to the popularity of *Populus* as an efficient model organism for forest biochemistry and genetics (Cervera *et al.,* 2001).

As a result of relatively higher rates of outcrossing and with pollen and seed flow occurring across wide geographic distributions, European aspen displays high levels of genetic diversity (Stevens *et al.*. 1999, Yeh *et al.* 1995 and Imbert and Lefèvre 2003). Recent advancements in sequencing and high-throughput genotyping have brought a fundamental change in genomic research (Wullschleger *et al* 2012). Analysis of whole genome expression data revealed that a transcript and gene expression level varies under different conditions (Kim *et al.,* 2012) and that there is extensive natural variation in the expression response both within and among *Populus* species (Street *et al.,* 2006). Studies of natural variation within a systems genetics or genetical genomics context can provide an additional means of identifying the mechanism (s) and polymorphisms underlying gene expression variation among individuals of a population. In order to advance such studies using *P. tremula*. The Umeå Plant Science Centre has initiated the 'aspen genome project'. In common with most current genome projects, the sequencing of aspen genome was performed using NGS methods (Shendure J and Hanlee J. 2008). However, the resultant assembly remains highly fragmented and assembled scaffolds are unordered. To facilitate ordering of scaffolds along chromosome a genetic map is required and assessing the most suitable NGS-based method of identifying molecular markers for map construction is the focus of this study.

The emergence and development of Next Generation Sequencing (NGS) technologies has enhanced almost all the biological disciplines making use of DNA sequence data in recent years (Gydle 2011 and Egan *et al.,* 2012). NGS techniques have made sequencing faster and easier and they typically generate enormous volumes of data (1 billion short reads per instrument run) of data at a relatively cheap price in per base pair terms compared to

traditional Sanger sequencing (Sanger *et al.*, 1977, Qiu *et al.*, 2010, Metzker 2010, Bao *et al.*, 2011 and Luca *et al.*, 2012). These new techniques have been applied to sequence numerous genomes ranging from small prokaryotic and virus genomes to those of eukaryotic animals and plants, many of which have large and complex genome composition. In contrast to microarrays, which can only profile expression of gene included during the array design, RNA-Sequencing profiles all expressed RNA captured by the selected sequencing library preparation method used (for example all polyA-RNA), which further allows finding unknown genes and splice variants of known genes (Luca *et al.*, 2012)

As a result NGS technology is replacing microarrays in different research applications. RNA-Seq does not require genome annotation for prior probe selection; however it does pose novel algorithmic and logistic challenges. Relatively lengthy procedures are required for current wet-lab RNA-Seq strategies, therefore NGS has been preferred in the projects involving transcript discovery in non-model organisms (Sacha *et al.*, 2010). Assembling NGS genomic data usually results in highly fragmented assemblies consisting of many thousands of contigs (Figure 1) with assembly of genes using RNA-Seq data is challenging due to the variable expression levels of genes and the presence of common sequence domains and gene families. NGS technologies have also improved gene annotation and identification of RNA alternative splicing events in comparison to classical application of DNA sequencing in genome resequencing projects and SNPs discovery (Bao *et al.*, 2011).

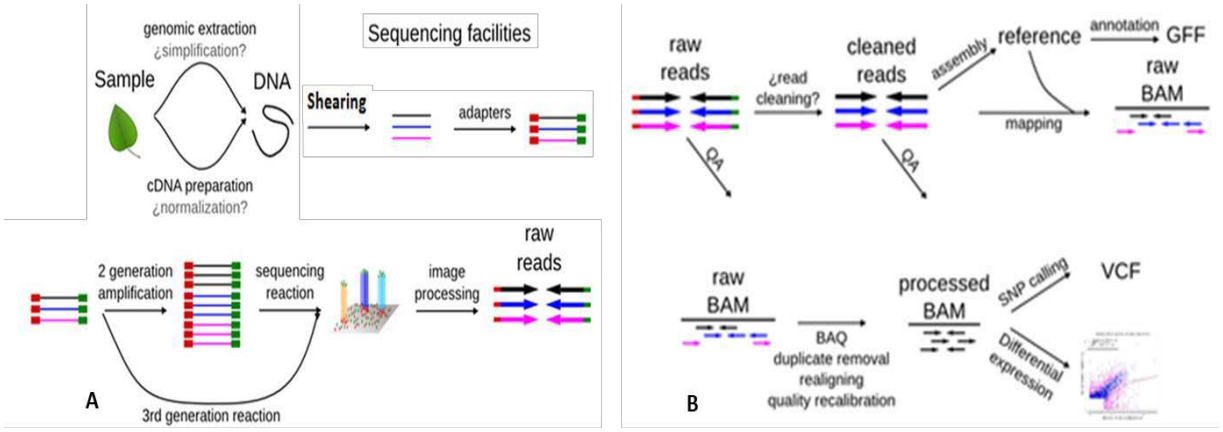


Figure 1: Next Generation Sequencing workflow. An example of NGS workflow is shown for the case of a non-model organism with no reference genome available, and therefore involves building a reference transcript by *de novo* assembly, mapping of raw reads to that assembly and finally SNP Calling and/or expression analysis.

Reduced representation library (RRL) construction and sequencing is method used to produce a comprehensive subset of the whole genome using restriction enzymes as a means of fragmenting the genome. As a result of genome fragmentation being achieved by the action of

restriction enzymes, almost all fragments produced will be in common among individuals of the same species with the only exceptions being those cases where a polymorphism is located within the enzyme restriction recognition site. Such RRLs represent a tractable representative subset of the genome for sequencing while still allowing the identification of extremely dense genetic markers for use in map construction (Young *et al.,* 2012). This technique was originally used for Single Nucleotide Polymorphism (SNP) discovery based on Sanger sequencing methods (Altshuler *et al.,* 2000 and Young *et al.,* 2010) where it was used to identify the SNPs from human samples (Luca *et al.,* 2012). More recently it has been applied using NGS methods in a range of species, but particularly so in non-model organisms (Van Tassell *et al.,* 2008 and Hyten *et al.,* 2012). Briefly, RRL sequencing involves re-sampling the same subset of the genome from several individuals and comparing the sequences obtained using an efficient SNP detection method. This analysis involves aligning sequences to form a consensus assembly of each represented restriction fragment and subsequently realigning the sequencing reads from the sampled individuals to that consensus to allow SNP detection (Altshuler *et al.,* 2000).

Next generation sequencing techniques typically involve randomly fragmenting genomic DNA and subsequent sequencing of the individual fragments (this is referred to as the shotgun approach). Fragmentation is required due to restrictions on the sequence length that can be produced, which in the case of NGS methods currently ranges from 50 bp to ~1Kbp, although longer read technologies are becoming available. Due to the requirement of fragment genomic DNA prior to sequencing, after obtaining the sequences of those fragments, the original non-fragmented (*i.e.* contiguous) sequenced must be reconstructed, which is achieved by application of genome assembly algorithms implemented as software tools. There are two distinct situations for performing an assembly, depending on whether a suitable reference genome is available to allow re-assembly on the basis of read alignments to that reference where no reference genome is available, in which case a *de novo* assembly must be produced.

Two basic approaches are used in algorithms for short reads assemblies *i.e.* "overlapped graphs" and "de Bruijn graphs" (De Bruijn *et al.,* 1946). In the case of the overlap approach, all sequence reads are pairwise aligned to identify the overlaps, a graph walking approach is then used to identify links between reads on the basis of these overlaps and, finally, a consensus sequence within regions of overlap is produced (the Overlap, Layout, Consensus or OLC approach).

In contrast to the OLC approach, the de Bruijn graph methods first breaks sequencing reads down into all possible Kmers (short sections of sequence) and links between those Kmers are identified followed by use of graph construction to reconstruct the longest possible stretches of contiguous sequence from those overlaps. This approach is required with NGS data primarily due to the fact that the use of Kmers reduces the amount of memory required by the assembly algorithm as well as reducing the number of potential overlaps that must be considered, which would be a significant problem if OLC methods were applied to the billions of reads generated by current NGS technologies.

Sanger sequencing based on cloning of the genes into a transformable vector, which is then multiplied whereas NGS methods involve fragmentation of DNA into thousands of small fragments, which are identified with the help of specific adapters in a huge population of small fragments. These fragments are then amplified in a multistep process during sequencing reactions to generate reads clusters and finally a population of raw reads (Figure 1-A).

The above described features of NGS methods are equally applicable to the sequencing of cDNA to profile gene expression levels (*i.e.* RNA-Seq), making it economical to produce large amounts of transcriptomic data, which can further provide information about expressed transcripts and complete and contiguous mRNA (Grabherr *et al.,* 2011, Simpson. *et al.* 2009, Trapnell, C. *et al.,* 2012 and Guttmann *et al.,* 2010). Reconstruction of full-length transcripts from short reads presents some challenges (Haas *et al.,* 2010), which have been addressed through computational solutions utilizing the de Bruijn graph assembly approach, which is the bottleneck for several whole genome assembly programs (Zerbino and Velvet 2008). One such algorithmic implementation is Trinity (Grabherr *et al.,* 2011), which efficiently reconstructs a large fraction of transcripts from duplicated genes and alternatively spliced isoforms, providing full-length recovery of transcripts with higher expression levels (Grabherr *et al.,* 2011).

Mapping (also referred to as alignment) reads against reference sequence assembly (either genome or transcriptome) is a key step in the analysis of NGS data (Horner *et al.,* 2009 and Bao *et al.,* 2011). A large number of NGS projects start with a reference genome, where positions of the reads must be determined through mapping. Mapping also needs to consider the amount of data and different characteristic error profiles produced from different platform version of the various NGS technologies (Trapnell *et al.,* 2009). For successful alignment, the most reliable reference sequence possible is needed, as assembly errors will prevent accurate alignment. Alignment tools are based on algorithms, which can be used to obtain maximum

information from sequencing data. Reads must be mapped with minimum gaps in the alignment in order to reduce the possibility miss-placed reads while allowing for the fact that technical sequencing errors do occur (Trapnell *et al.,* 2009). Mapping is particularly challenging in the case of RNA-Seq, where a mature mRNA is converted into cDNA and sequenced to enable the identification of previously unknown genes and alternative spliced variants and where alignments are performed with large gaps due to presence of introns (Trapnell *et al.,* 2009). These challenges are not new and can be handled by many programs offering spliced and unspliced alignments (Bozdag *et al.,* 2009). DNA sequencers from different companies (Illumina, Helicos, ABI and Roche of Basel) generate millions to billions of reads per run and complete assays may involve many runs. It is now possible to map billions of reads to a reference sequence, while a large and expensive computer grid could potentially map the reads in a few days using traditional alignment algorithms, such as BLAT or BLAST, mapping reads from Chip-Seq or RNA-Seq data would require thousands of central processing unit (CPU) hours using such tools. Since these grids are not accessible by everyone, such resources are in limited supply and the computing cost of sequence-based analysis is relatively high, a new generation of alignment programs has been generated. These programs map hundreds of millions of short reads on a single desktop computer (Trapnell *et al.,* 2009). These tools include user definable parameters to account for features such as error characteristics, expected variation compared to the available reference, expected maximum intron size etc. (Trapnell *et al.,* 2009).

The method used to create an addressable index of either the reference genome or reads to be aligned has categorized these tools (Horner *et al.,* 2009, Li *et al.,* 2009 and Bao *et al.,* 2011). Some software implementations, such as Eland, Cloud-Burst, MAQ, ZOOM, SeqMap, SHRiMP, and RMAP, are based on constructing hash tables for short reads, then aligning them to the original genome sequences. The memory usage of these programs depends on the number of reads to be processed. Another category of software (BWA, Bowtie, MOM, BFAST, PASS and SOAP etc.) includes those implementations that index genomic sequences. It is easy to parallelize this class of software to utilize multithreading of single CPUs where available or to divide the computations across multiple CPUs (Bao *et al.,* 2011). The mapping programs Maq (Li *et al.,* 2008) and Bowtie (Langmead *et al.,* 2009) are based on indexing the reference genome to enhance mapping speed. Maq relies on a simple and effective strategy called spaced seed indexing (Li *et al.,* 2012). According to this strategy, "seeds" are generated by splitting the reads into four equal length segments, which are then

aligned against the reference genome with alignments then being extended across the remaining section of the read. Bowtie is an ultra-fast, memory efficient short read aligner, capable of aligning large sets of short DNA sequences (reads) to relatively large genomes (up to a maximum of ~3 Gbp). It uses the Burrows-Wheeler Transform for genome indexing (maximum genome sizes of 2.2 Gbp for unpaired and 2.9 Gbp for paired-end alignment) and for minimizing memory footprint. It outputs alignments in the now standard Sequence Alignment Map (SAM) format, which interoperates with numerous other tools (Langmead *et al.,* 2009). Bowtie creates a permanent index, which may be re-used across alignment runs. Furthermore, Bowtie uses standard FASTQ and FASTA input formats and is provided with a conversion program to allow Bowtie outputs to be used with Maq's consensus generator and SNP caller (Langmead *et al.,* 2009).

A genetic map is a representation of a genome where the recombination frequencies between polymorphic loci are used to positioned markers relative to one another (Barbazuk *et al.,* 2005). Until recently microsatellite markers had been most commonly used for genetic maps, although there are several different types of genetic markers available. More recently, genetic maps with very high marker density (and therefore map resolution) are being constructed using single nucleotide polymorphisms markers. Linkage maps constitute the framework for using genetic markers in marker-assisted selection (MAS) breeding programs (Mazur and Tingey 1995). Although there are a number of genetic maps available for different *Populus* population, no sequence-based genetic map is currently available for *P. tremula*. A number of SSR microsatellite markers known to produce amplification products across many *Populus* species, including *P. tremula,* are available however there are too few to allow construction of a high density genetic map, as is required for facilitating placement of assembled scaffolds along chromosomes from a draft genome assembly. As such the identification of high-density SNPs markers throughout the genome represents an excellent option for generating such a map.

On account of the high level of heterozygosity and gene duplication, aspen presents challenges for mapping and sequencing efforts (Kelleher *et al.,* 2007), where high level of heterozygosity may cause independent assembly of haplotypes in hyper variable genomic regions (Kelleher *et al.,* 2007). Similarly, mis-assembly may also occur due to genomic sequence with high sequence similarity at multiple locations within the genome. Aspen is a dioeciously out-crossing species with relatively higher level of gene flow. On account of its

wind-pollinated nature, the haplotype diversity rates are highly increased *i.e.* 2.6 polymorphisms (SNPs)/ Kbp (Tuskan *et al.,* 2006 and Kelleher *et al.,* 2007).

The aim of this project was to assess two alternative approaches for identifying high-density polymorphic genetic markers between the *P. tremula* parents of an F_1 population to allow construction of a genetic map. We found that DNA of adequate quality to allow RRL construction could only be obtained using long and manually intensive DNA extractions protocols rendering this approach infeasible for application within the F_1 population. We therefore focused our efforts on the use of existing RNA-Seq data that was used to generate a *de novo* transcript reference that we subsequently used for read alignment and SNP detection using RNA-Seq data from the two parental genotypes.

2. Materials and methods

2.1. Plant material and DNA Extraction:

Different leaf samples from European aspen were used in order to obtain genomic DNA. Initially, freeze dried aspen leaves (collected from an experimental field located at Säver, north western part of Umeå) were used for the male parent (labeled as 229.1) and female parent (labeled as 349.2), stored in -80 ^{o}C for DNA extraction. Preserved samples leaves (20 mg) in liquid nitrogen were transferred into an eppendorf tube containing the hot extraction buffer. Extraction buffer (4 ml of 0.5 M EDTA, 10 ml of 1M Tris-Cl, 28 ml of 5 M NaCl, and 2% CTAB (2g)) and samples were mixed together to eliminate the clumps. Samples were shacked in water bath at 65 ^{o}C and were incubated for 25 minutes inverting the tubes 2-3 minutes during the incubation. Eppendorf tubes were cooled down at room temperature and an equal volume of chloroform: isoamyl alcohol (24:1) was added and the tubes were inverted 20-25 times. Tubes were centrifuged at 11000 rpm for 15 minutes. The upper phase from the tubes was transferred into fresh tubes and 0.5 volume of M NaCl and 2 volume of 95% ethanol (stored at -20 ^{o}C) were added. Tubes were inverted several times before putting in the freezer at -20 ^{o}C for 20-30 minutes. Samples were centrifuged at 14000 rpm for 10 minutes at 4 ^{o}C. Supernatants were removed and 700 µl of 80% ethanol was added. Samples were again centrifuged at 14000 rpm for 5-7 minutes. Ethanol was removed and tubes were let opened to be dried. Finally, 20-50 µl of H_2O was added to dissolve the pellet. Later, for a comparative analysis of restriction products, three week old leaves from *Arabidopsis thaliana* wild type Columbia-0 ecotype were also used for DNA extraction. For Arabidopsis DNA extraction, a miniprep protocol from QIAGEN DNA extraction kit was used according to manufacturer's instruction (DNeasy Plant Handbook 07/2006, Umeå, Sweden). The extraction buffers required for DNA extraction were supplied with the extraction kit.

The quantity and quality of DNA for both Arabidopsis and aspen was assessed through QUBIT (Invitrogen Qubit+ds DNA BR Assay kit). The Qubit 2.0 Fluorometer is a benchtop fluorometer for the quantification of DNA, RNA and proteins, by using a highly sensitive and accurate fluorescence-based Qubit quantification assay. Effects of contaminants have been minimized through the use of state-of-the-art dyes selective for dsDNA, RNA and proteins, along with the latest illumination and detection technologies used in the Qubit 2.0 Fluorometer, which provides the highest sensitivity at the cost of as little as 1µl of the sample (Appendix 1, Table 4).

2.2. Reduced Representation Library Preparation and High Throughput Sequencing:

Identification of genetic markers was studied in European aspen (*Populus tremula*) grown in the Umeå region. Restriction analysis of genomic DNA through endonucleases helps to create reduced representation libraries, based on short segments of the fragmented DNA (300-600 bp). DNA samples were digested with fast digesting restriction enzymes at 37 ^{o}C for 15-20 minutes. In order to compare the efficiency of DNA extraction protocols and restriction enzymes, seeds from *Arabidopsis thaliana* ecotype Col-0 were also used. During the wet lab experiments a set of restriction endonucleases enzymes was applied to restrict the aspen genome (Figure 4). A set of 19 widely used restriction enzymes were used for this study (Table 1). The reason behind getting the short fragment length reads was the capacity of sequencing tools to sequence these fragments. DNA fragments, falling beyond the range of 300-600 bp cannot be read by any of these methods. These enzymes were applied individually and in combination as well with other enzymes to obtain the expected size (bp) of restriction fragments. The samples were run on 1% agarose gels after RNAse treatment. A total volume of 30 µl was loaded on the gel.

Table 1: List of Endonucleases enzymes applied during restriction analyses. It represents the details of 19 restriction enzymes used for the restriction of aspen DNA. These enzymes were used separately and in combinations with other enzymes.

Sr. No	Restriction Enzyme	Source	Recognition site	Cut
1	*Acc65I*	*Acinetobacter calcoaceticus* 65	5' GGTACC 3' CCATGG	5' ---G GTACC--- 3' 3' ---CCATG G--- 5'
2	*BamHI*	*Bacillus amyloliquefaciens* H	5' GGATCC 3' CCTAGG	5' ---G GATCC--- 3' 3' ---CCTAG G--- 5'
3	*BseNI*	*Bacillus* sp. N	5' ACTGG 3' TGACC	5' ---ACTGGN --- 3' 3' ---TGAC CN--- 5'
4	*BsrBI*	*Bacillus stearothermophilus* CPW193	5' CCGCTC 3' GGCGAG	5' ---CCG CTC--- 3' 3' ---GGC GAG--- 5'
5	*ClaI*	*Caryophanon latum* L	5' ATCGAT 3' TAGCTA	5' ---AT CGAT--- 3' 3' ---TAGC TA--- 5'
6	*EcoRI*	*Escherichia coli* RY13	5' GAATTC 3' CTTAAG	5' ---G AATTC--- 3' 3' ---CTTAA G--- 5'
7	*EcoRV*	*Escherichia coli* J62 pLG74	5' GATATC 3' CTATAG	5' ---GAT ATC--- 3' 3' ---CTA TAG--- 5'
8	*HaeIII*	*Haemophilus aegypticus*	5' GGCC 3' CCGG	5' ---GG CC--- 3' 3' ---CC GG--- 5'

9	*HindIII*	*Haemophilus influenzae* Rd	5' AAGCTT 3' TTCGAA	5' ---A AGCTT--- 3' 3' ---TTCGA A--- 5'
10	*HpaII*	*Haemophilus parainfluenzae*	5' CCGG 3' GGCC	5' ---C CGG--- 3' 3' ---GGC C--- 5'
11	*MseI*	*Micrococcus* sp.	5' TTAA 3' AATT	5' ---T TAA--- 3' 3' ---AAT T--- 5'
12	*NotI*	*Nocardia otitidis-caviarum*	5' GCGGCCGC 3' CGCCGGCG	5' ---GC GGCCGC--- 3' 3' ---CGCCGG CG--- 5'
13	*NsiI*	*Neisseria sicca*	5' ATGCAT 3' TACGTA	5' ---ATGCA T--- 3' 3' ---T ACGTA--- 5'
14	*PstI*	*Providencia stuartii* 164	5' CTGCAG 3' GACGTC	5' ---CTGCA G--- 3' 3' ---G ACGTC--- 5'
15	*Sau3AI*	*Staphylococcus aureus* 3A	5' GATC 3' CTAG	5' --- GATC--- 3' 3' ---CTAG --- 5'
16	*StyI*	*Salmonella typhi*	5' CCWWGG 3' GGWWCC	5' ---C CWWGG--- 3' 3' ---GGWWC C--- 5'
17	*SwaI*	*Staphylococcus warneri*	5' ATTTAAAT 3' TAAATTTA	5' ---ATTT AAAT--- 3' 3' ---TAAA TTTA--- 5'
18	*Xbal*	*Xanthomonas badrii*	5' TCTAGA 3' AGATCT	5' ---T CTAGA--- 3' 3' ---AGATC T--- 5'
19	*Xhol*	*Xanthomonas holcicola*	5' CTCGAG 3' GAGCTC	5' ---C TCGAG--- 3' 3' ---GAGCT C--- 5'

2.3. *In silico* Restriction analyses:

In silico restriction digestion of the aspen genomic DNA was performed with two restriction enzymes *i.e. PstI* and *MseI* to obtain the small fragments of DNA. Aspen genome sequence was downloaded from an open source database i.e. “www.phytazome.org”. To restrict the aspen genome into several fragments, a software program tool called "emboss restrict" (http://helixweb.nih.gov/emboss/html/restrict.html) was used. This program works on the basis of some parameters, which should be selected before executing the program. Among these parameters, some are “minimum cuts per enzyme”, “maximum cuts per enzyme”, “blunt and sticky ends allowed” etc. (Figure 2). R-program language was used to write the script to draw frequency distribution histograms afterward (Figure 5).

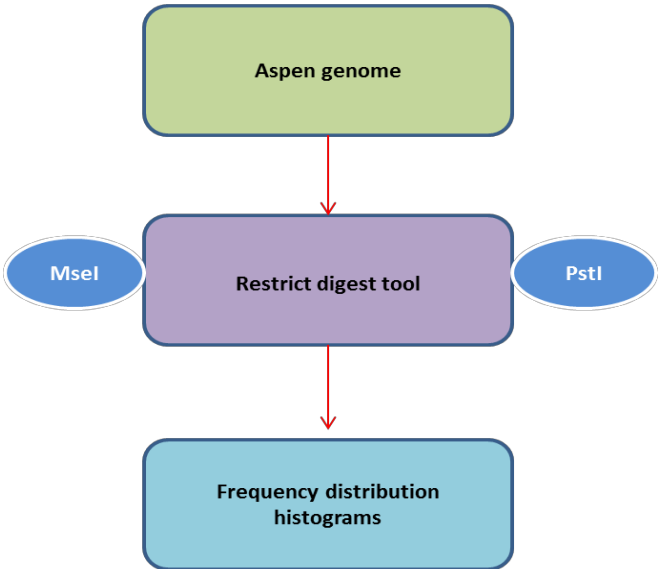


Figure 2: Workflow for the *in silico* restriction digestion of aspen genome by two enzymes (*MseI* and *PstI*). It shows the steps involved in *in silico* restriction digestion. Aspen genome is restricted by two restriction enzymes i.e. *MseI* and *PstI*. In restrict digest tool, certain parameters are set according to the nature of restriction enzyme.

To accomplish this study, two different approaches were assessed. Firstly, applicability of utilizing a reduced representation sequencing library approach in the parents of an F_1 intraspecific *P. tremula* population was assessed. Secondly, the use of existing gene expression data that had been generated using RNA-Sequencing (RNA-Seq) to generate a *de novo* reference transcript assembly for one of the parents of the *P. tremula* F_1 population was assessed.

2.4. Application of Computational tools and SNP Calling:

To download all the genome sequences, different databases were used i.e. “Galaxy” (http://galaxy.popgenie.org:8080), “PopGenIE” (www.popgenie.org) and “The Arabidopsis Information Resource” (TAIR). All the software tools were run on galaxy and high memory servers through Linux operating system. Galaxy is an open and web-based server/database and a tool kit that encompasses many NGS tools for the data analysis (Sjödin *et al.*, 2009). First of all, RNA-Seq reads for both samples were screened for good quality reads through FastQC. It facilitates to keep only the good quality raw sequences produced through high-throughput sequencing pipelines. FastQC interprets the raw sequence data in terms of graphs and tables to assess the data, which are then transformed into an HTML report. It aims to provide some quality control checks on raw sequence data through measuring quality scores across all the bases (Illumina 1.5 encoding), per sequence quality scores, per base sequence

content, per base GC content, per base N content, sequence length distribution, sequence duplication level, overrepresented sequences and Kmer contents. FastQC also generates a basic statistics table mainly representing total sequences, filtered sequences, sequence length and GC percentage. Sequence reads for both samples were then groomed through FastQ groomer, which offers several conversion options relating to FastQ format. During the conversion between Solexa and other formats, quality scores are mapped between Solexa and Phred scales (Cock *et al.,* 2009). After trimming the raw sequence reads, Illumina sequences (in FastQ format) for the samples 229.1 (male parent) were used for *de novo* assembly by Trinity (Grabherr *et al.,* 2011). It represents a novel method for the efficient and robust de novo method for reconstruction of reference transcriptome from RNA-Seq data. Trinity includes three independent software modules, which are Inchworm, Chrysalis and Butterfly to process large volume of RNA-Seq reads. It distributes the sequence data into many individual de Bruijn graphs, each representing a transcription complexity at a given gene or locus. After *de novo* assembly of reference transcript, both samples of RNA-Seq reads were aligned against the reference transcript using "Burrows-Wheeler Aligner", BWA (Li H. and Durbin R. 2012). The qualities of output files from BWA were further screened to avoid bad quality reads. The BWA output files were then processed to trim poor quality sequence reads by using FastQC groomer and Flagstat. Flagstat is used to produce simple statistics by summarizing the flags produced within BAM files based on SAM Tools, showing percentage of mapped reads, properly paired and singletons reads percentage (Table 3).

The sequence reads for both samples (229.1 and 349.2) were aligned separately to reference transcript. The sequence reads were indexed and mapped using BWA (Li H. and Durbin R. 2012). The BAM-files were then used for the Genome Analysis Toolkit (GATK) (Figure 3). Trinity transcriptome data was also aligned against the *P.trichocarpa* genome with the help of a Genomic Mapping and Alignment Program (GMAP). SNP Calling was performed in two steps; the primary data for raw SNPs were collected by a program tool i.e. GATK-Analysis**.** GATK (McKenna *et al.,* 2010) involves base quality score recalibration, indel realignment, duplicate removal and performed SNP and INDEL discovery. It also performs genotyping across the samples with the help of standard hard filtering parameters or variant quality score recalibration (DePristo *et al.,* 2011).

After executing the preliminary steps to screen raw SNPs and checking the quality of an output file 'BAM' (a compressed binary version of Sequence Alignment Map (SAM). Unified Genotyper was run to generate a VCF (Variant Call Format) file. After mapping trinity

transcripts data to *P. trichocarpa* by using GMAP, an output (General Feature Format) 'GFF' file was used to identify/locate the entire transcripts only for chromosome number 19 (being sex chromosome) in a VCF file to identify (by using intersectBED) the transcript sequences for primer designing.

Finally, the putative SNPs were filtered based on SNP quality (QUAL) sequencing depth (DP), allelic frequency (AF) and genotypic information (GT). SNP quality (QUAL) and sequencing depth (DP) values were compared through a function "comparisonplot" in R program language. The initial output file containing all the SNPs was processed to draw a subset of high quality SNPs by Linux and R scripts.

2.5 Identification of SNPs in genome browser: Integrative Genome Viewer (IGV)

Single Nucleotide Polymorphisms (SNPs) genetic markers were also identified in the genome browser "Integrative Genome Viewer" (IGV). SNPs with their particular transcript IDs were observed for their homozygous or heterozygous nature. A list of all validated SNPs was annotated by using a software tool "snpEff", which predicts the effect of SNPs located within coding regions and classifies the genomic context of SNPs (intronic, exonic, intergenic) given a supplied genome annotation.

A specific workflow was generated to perform all the downstream analyses. This workflow used alignments of the reads to *P. trichocarpa* for identifying the putative SNPs between sequence reads for both samples (Figure 3).

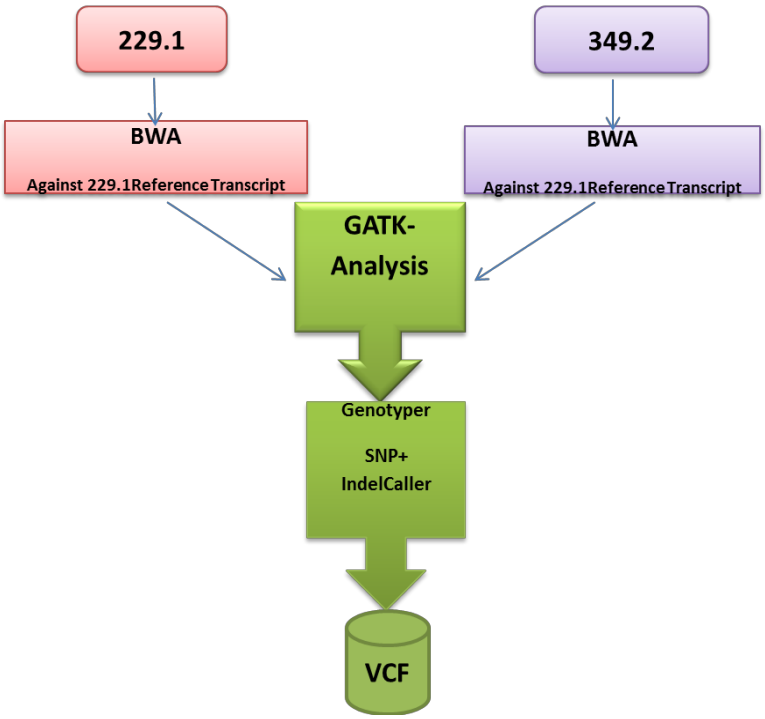


Figure 3: workflow for different tools to run on galaxy. It represents the scheme of mapping the reads and SNP-Calling through a local toolkit 'Galaxy'. Sequence reads for both samples were mapped using BWA and SNPs were called by using GATK-Analysis toolkit.

3. Results

To accomplish this study, two different approaches were assessed. Firstly, applicability of utilizing a reduced representation sequencing library approach in the parents of an F_1 intraspecific *P. tremula* population was assessed. In the second alternative approach, the use of existing gene expression data that had been generated using RNA-Sequencing (RNA-Seq) to generate a *de novo* reference transcript assembly for one of the parents of the *P. tremula* F_1 population was applied to facilitate construct of a genetic map of SNPs genetic markers on chromosome 19 of *P. tremula*.

3.1. Assessing the potential of RRL preparation in *P. tremula*

Freeze dried leaves were used to extract aspen DNA. Initially, after repeated DNA extraction experiments with different extraction protocols, expected quality and/or quantity of DNA was not obtained both for aspen and Arabidopsis leaves. The DNA concentration measured for female parent was observed as 2-4 ng/µl. Similarly, DNA concentration measured for the male parent was also insufficient (6-10.5 ng/µl). However after optimizing the extraction protocol for aspen DNA extraction, better quality DNA was extracted (approximately 500 ng/µl) whereas for Arabidopsis, a quantity of approximately 400 ng/µl of DNA was obtained after extraction through QIAGEN DNeasy Mini prep extraction kit for restriction analysis.

3.2. Restriction analyses

Restriction analysis of genomic DNA from aspen leaves was performed to restrict the DNA into short fragments. Aspen genomic DNA was restricted with 19 different endonucleases enzymes (Table 1). After repeated DNA extractions and restriction digestion trials by different enzymes, expected DNA fragments were not obtained (Figure 10-12, Appendix 2). In order to screen out the most effective restriction enzymes activity, a single enzyme (*EcoRI*) was used to test both the DNA quality and enzyme's activity (Figure 10-A). The number of enzymes was increased from 1 to 19 with successive investigation hits but none of these restriction enzymes clearly showed the expected fragments of DNA both from aspen and Arabidopsis (Figure 10-12, Appendix 2).

Aspen and Arabidopsis genomic DNAs were also restricted with the help of *EcoRI* and *PstI* enzymes used as single enzyme per lane (Figure 10, Appendix 2). Aspen DNA restricted with *EcoRI* show a smear around 10000-20000 bp, whereas Arabidopsis DNA restricted with *PstI* produced a smear in same area i.e. around 20000 bp.

Aspen and Arabidopsis genomic DNAs were also restricted with the help of other enzymes used as single and in different combinations with each other (Figure 11-A, B, Appendix 2) with other restriction enzymes. Restriction enzymes applied on aspen DNA e.g. *EcoRI*, *PstI*, *BamHI* and *HindIII* produced a smear around 20000 bp regions (Figure 11, Appendix 2). The DNA from Arabidopsis was also restricted by the combination of different enzymes but none of them show clear restriction of Arabidopsis genomic DNA.

Apart from the usual restriction digestion of aspen and Arabidopsis DNA with different restriction enzymes, the varaible quantities of aspen DNA were also applied for the restriction analysis (Figure 12, Appendix 2). *HpaII* was used with different quantities i.e. 2µl and 3µl (Figure 12-B) in order to obtain desired restriction of aspen DNA. After applying different restriction enzymes with variable quantities, all the enzymes produced smears around 7000-20000 bp fragment size.

Finally, by using two different DNA extraction protocols, the expected restriction products for both both *Populus* and Arabidopsis genomic DNA were obtained (Figure 4). *Populus* genomiv DNA was digested into different fragment sizes depending on the restriction sites available on restriction enzymes. A population of many small fragments was observed ranging between 75-400 bp by using *MseI* both for *Populus* and Arabidopsis DNA, whearas aa population of relatively larger fragments of DNA was observed at the region ranging between 3000-10000 bp fragments both for *Populus* and Arabidopsis DNA by using *PstI* restriction enzyme. Similarly, same digestion pattern was observed when both of these restriction enzymes used together in combination (Figure 4).

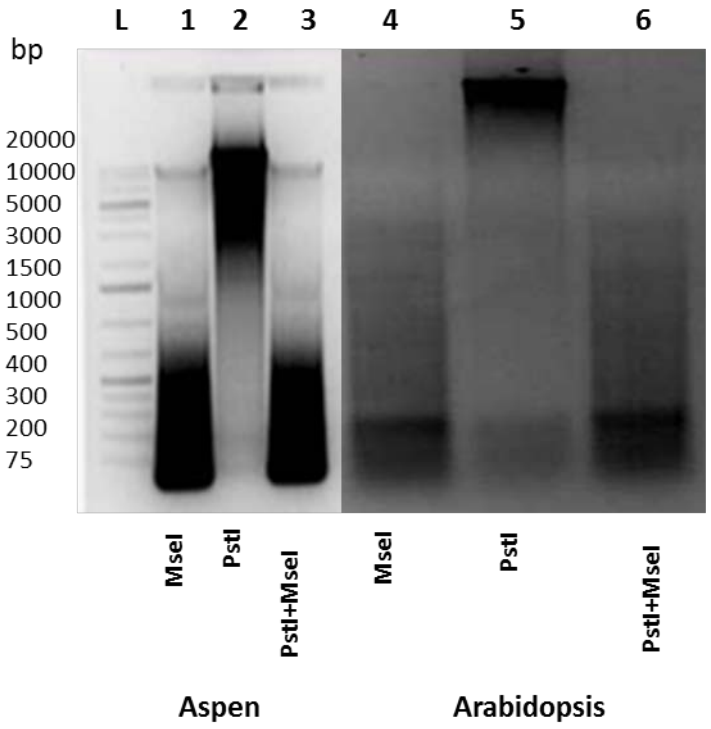


Figure 4: Restriction analyses of aspen DNA by different individual restriction enzymes. Restriction analysis by restriction enzymes ***MseI*** and ***PstI,*** where **L**= ladder, lane 1, 2 and 3 shows restriction of aspen DNA with *MseI*, *PstI* and mixture of both enzymes respectively, similarly lane 4, 5 and 6 represents restriction digestion of Arabidopsis DNA by same enzymes with same order.

3.3. *In-silico* Restriction analyses

In-silico restriction digestion of genomic DNA of *Populus trichocarpa* tree was performed by using an open source software program called "Restrict" with two restriction enzymes i.e. *PstI* and *MseI*. The parameters required for running this tool were set as "minimum cuts per enzyme=1", "maximum cuts per enzyme=2000000000," blunt and sticky ends allowed" etc. The restrict program was run on the entire genome of *Populus trichocarpa* containing total 19 number of chromosomes and 1427 number of scaffolds. This software produced 4748637 restriction products by the *MseI* enzyme, whereas a total of 75939 restriction products were found by the enzyme *PstI* from the entire *P.trichocarpa* genome. Based on these scaffolds frequency distribution graphs were generated to present the data in readable form by writing an R script (Figure 5-A, B and C). Maximum number of DNA fragments were found to be of small size i.e. 0-100 bp (Figure 5-A) not only by individual enzymes but in combination also (Figure 5-C).

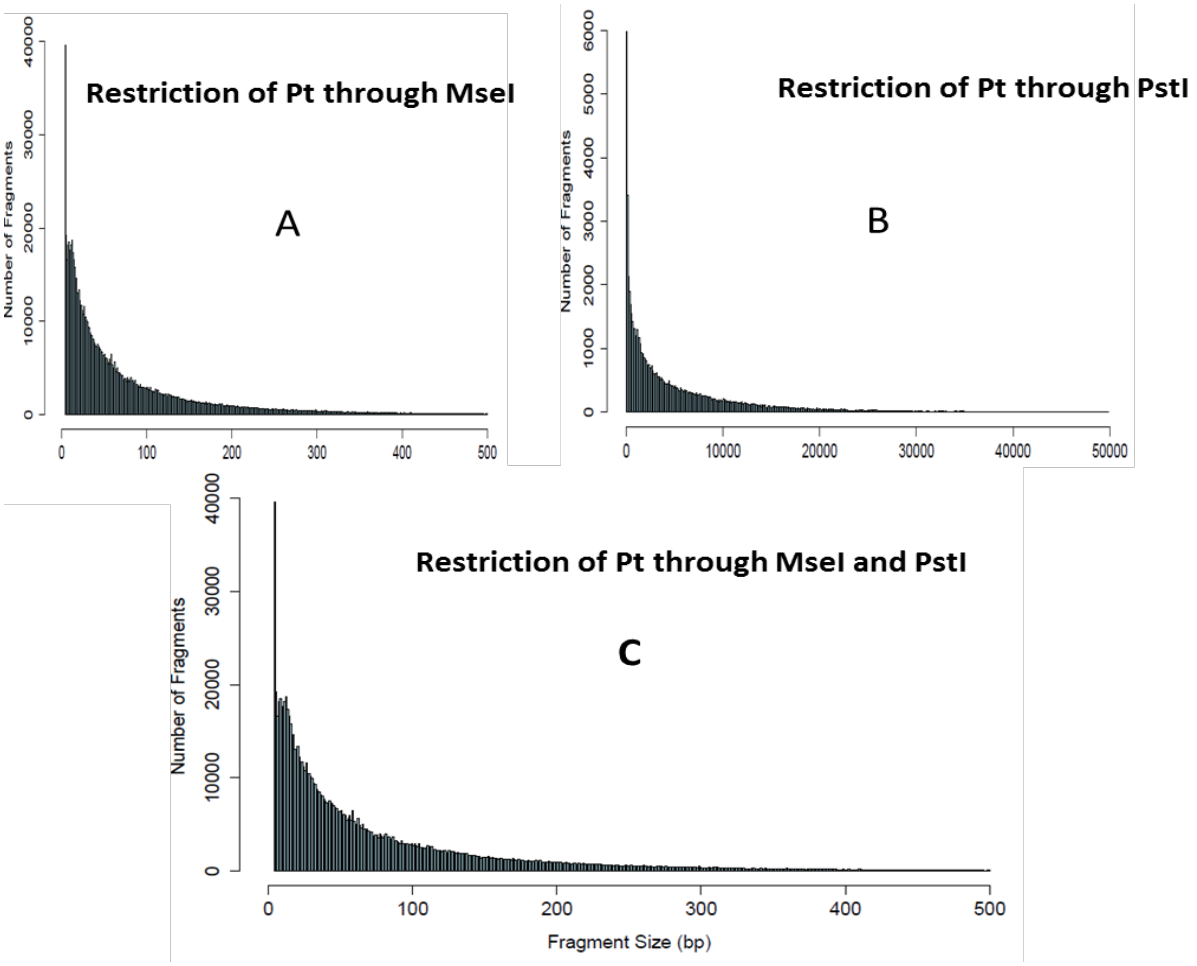


Figure 5: *In silico* restriction analyses of aspen DNA by *MseI* and *PstI* restriction enzymes through restrict.digest program. A represents the frequency distribution by *MseI* and **B** represents frequency distribution of *PstI*, whereas **C** represents the frequency distribution of aspen genome restricted with both enzymes (*MseI* and *PstI*) together.

3.4. Application of Computational tools and SNP Calling:

3.4.1. *De novo* assembly

Initially, to compare sequence reads for both of the samples, a reference transcript sequence was built by using a *de novo* assembler "Trinity". The initial transcript data was trimmed / filtered through a python script and finally, a total of 95854 transcripts were assembled together to build a reference transcript. These transcripts were then aligned to *P.trichocarpa* genome using GMAP, which resulted in a total of 94,762 aligned sequence reads and 1,092 unaligned sequence reads.

3.4.2. Galaxy

Extraction of *Poplulus* genomic DNA was not as successful as it was anticipated in the beginning of the project, which also produced unexpected restriction products for the

construction of reduced representation libraries. Therefore, all the computational analyses were then performed based on RNA-Seq data for both samples (229.1 and 349.2). We used our local server and tool kit “Galaxy” for different computational tools. The sequence reads for both male and female parental lines (229.1 and 349.2 respectively) were primarily trimmed by applying the cut adapters. To follow all the downstream analysis, first of all, both sequence reads were screened for quality control checks. Sequence reads for female sample (349.2) were, first screened for quality score for all bases (Illumina 1.5 encoding) (Figure 9). Mean sequence quality (Phred score) was measured as 37 covering 450000 reads. Total GC% contents were measured as 44 % and sequence duplication level was measured as more than 54 % with no overrepresented sequence and kmers (Table 2). Similarly, sequence reads for male parent (229.1) were also screened for quality checks and mean sequence quality (Phred score) was measured as 38 covering more than 450000 reads. Total GC % content for male (229.1) sequence reads were also measured as 44% and sequence duplication level was measured as more than 54 % with no overrepresented sequence or kmers (Table 2). These sequence reads were then aligned against 229.1 based trinity transcript data (reference sequence) with BWA by using the default parameters (Li H. and Durbin R. 2012). The output SAM files were converted to more compressed form i.e. BAM files by using SAM-Tools. The output files from BWA were further checked for the qualities by removing the bad quality sequence reads by using FastQC groomer and Flagstat. After aligning the 349.2 sequence reads by using BWA, a total of 86 % reads were aligned to the 229.1 reference sequence, of which approximately 77 % were properly paired and 4.55 were singletons, whereas 89 % of the total reads from male sample (229.1) were aligned against 229.1 reference sequence. Properly paired reads were measured as 80 % and 4 % of the aligned reads were found to be singletons (Table 3).

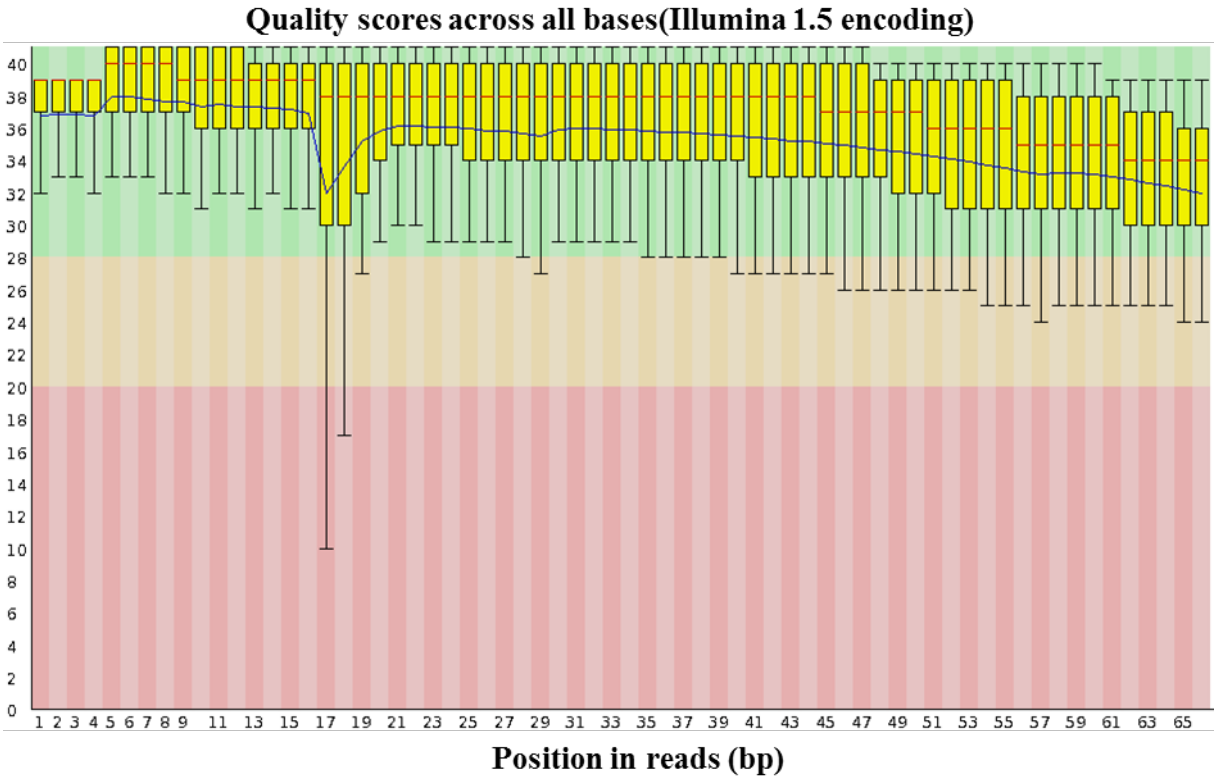


Figure 6: FastQC quality control over reads, showing per base sequence quality of reads. It represents the quality of sequence reads. Quality scores for all bases of the 349.2 sequence reads are shown. Most of the bases represent good quality scores whereas, there are some poor quality bases also i.e. base 17 and base 18.

Table 2: Information about the sequences of two data sets (229.1 and 349.2) after grooming/trimming the sequences. It shows the summary of FastQC reports with basic statistics about the both samples reads. Quality scales were measured based on Illumina 1.5 encoding scale. Both samples were observed with no overrepresented sequences and similar GC contents.

	229.1	**349.2**
Measure	**Value**	**Value**
Filename	dataset_229.1.dat	dataset_349.2.dat
File type	Conventional base calls	Conventional base calls
Encoding	Illumina 1.5	Illumina 1.5
Total Sequences	33330830	32199905
Filtered Sequences	0	0

Sequence length	27-66	27-66
%GC	44	44
Overrepresented sequences	0	0
Sequence duplication level	57%	54%

Table 3: Flagstat statistics on both samples reads (229.1 and 349.2). It represents the flagstat simple statistics on both samples. The numbers of reads aligned to the reference transcript were variable in number in both samples in each successive flagstat reports with minor differences.

Flagstat statistics	**229.1**	**349.2**
Total (QC-passed reads + QC-failed reads	66661660	64399810
Mapped	89.28%	86.30
properly paired	79.82	76.73
Singletons	4.15%	4.59

3.4.3. SNP Calling

SNP Calling was performed using "GATK-Analysis". The BWA output files (BAM files) for both parental sequence reads (229.1 and 349.2) were used as input files for GATK-Tools **(**Appendix 3, Figure 13**).** After mapping trinity transcripts data to *P. trichocarpa* by using GMAP, an output (General Feature Format) 'GFF' file and GATK output file 'VCF' for both types of sequences reads containing raw SNPs were compared to identify the putative homozygous SNPs.

In order to identify the raw SNPs, the primary data for raw SNPs were obtained by using GATK-Analysis toolkit**.** After executing the preliminary steps to screen out raw SNPs and checking the quality of BAM (a compressed version of Sequence Alignment Map (SAM)) files, Unified Genotyper was run to generate a 'Variant Call Format' (VCF) file. GMAP based GFF and Browser Extensible Data (BED) (VCF converted to BED format) files were used to read and locate the entire transcript IDs linked with chromosome number 19 only (sex chromosome) in VCF file to obtain (by using intersectBED) the sequences of putative SNPs

in the entire genome sequence. The final output file containing 8122 SNPs was obtained through writing Perl and R scripts. These SNPs were located in the entire *Populus trichocarpa* genome sequence.

3.4.4. SNP Filtering

Finally, in order to obtain a high quality subset of SNPs, all the SNPs from the initial raw SNP storing VCF file were filtered. SNP quality (QUAL) and sequencing depth (DP) values plotted through an R function called "comparionplot" (Figure 10). Taking log10 values both for QUAL and DP further helped in sub-setting the high quality SNPs. All the transcripts with SNP Quality, sequencing depth, allelic frequencies and genotypic information were extracted to filter a sub set of relatively higher quality SNPs (Appendix 4, Table 5).

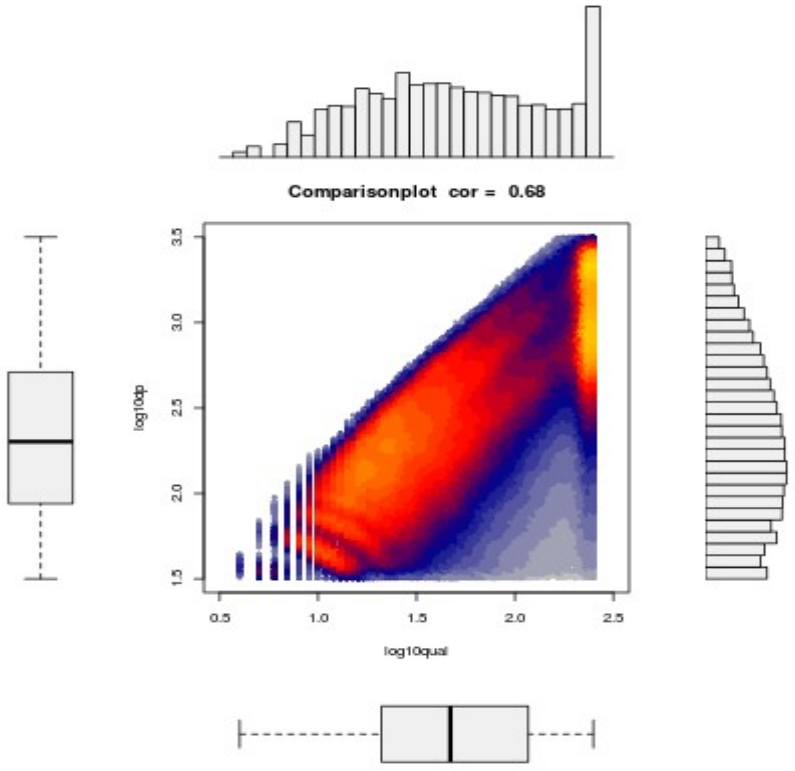


Figure 7: Comparisonplot between SNP quality (qual) and sequencing depth (dp) values. It represents the comparisonplot between log10 of SNP quality and sequencing depth values. Based on the red cloud on y-axis, a subset of SNPs was obtained after applying a threshold of 3.0-3.2 on dp values and 2.0-2.2 on qual values.

3.4.5. SNP location/position through Integrative Genomics Viewer (IGV)

The extracted pool of high quality SNPs was also identified through Integrative Genomics Browser (IGV) by aligning BAM files for both of the samples (229.1 and 349.2) and VCF file consisting of filtered set of 118 SNPs of higher quality, against 229.1 based reference

transcript. These high quality SNPs were identified with the help of their specific IDs and base pair position in the reference transcript. Single nucleotide base pair present in all the reads in each respective file was attributed as homozygous whereas, those nucleotide bases present in approximately half of the reads, were attributed as heterozygous SNPs.

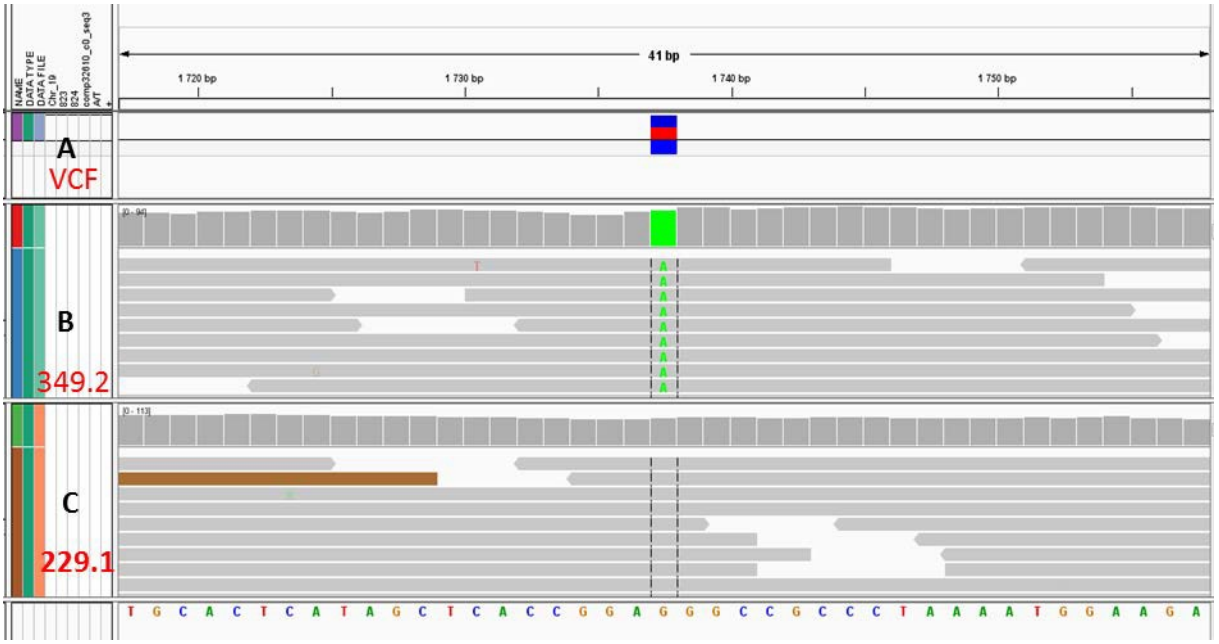


Figure 8: Identification/location of high quality SNPs in the Reference transcript through genome browser (IGV). A indicates a SNP from the VCF file containing extracted SNPs, **B** indicates homozygous SNP in female (349.2) parent and **C** indicates male parent (229.1) with no SNPs in the sequence reads against the 229.1 based reference transcript.

Finally, all the 118 SNPs were annotated to find their effects on genes. A tool (snpEff) that annotates and predicts the effects of variants on genes (such as amino acid changes) was used to see the synonymous or non-synonymous changes. Out of 118 a single (1) SNP was found to be homozygous, whereas remaining 117 SNPs were found as heterozygous SNPs and 98 % of filtered SNPs were observed as modifiers. Thirty-two (32 %) of the filtered SNPs were found in intergenic region. The final set of filtered SNPs was also analysed to measure the synonymous and non-synonymous effects of these SNPs, where all the SNPs present in coding regions, affected the genes responsible for protein encoding with 0.28 % non-synonymous and 0.85 % synonymous effects (Figure 9).

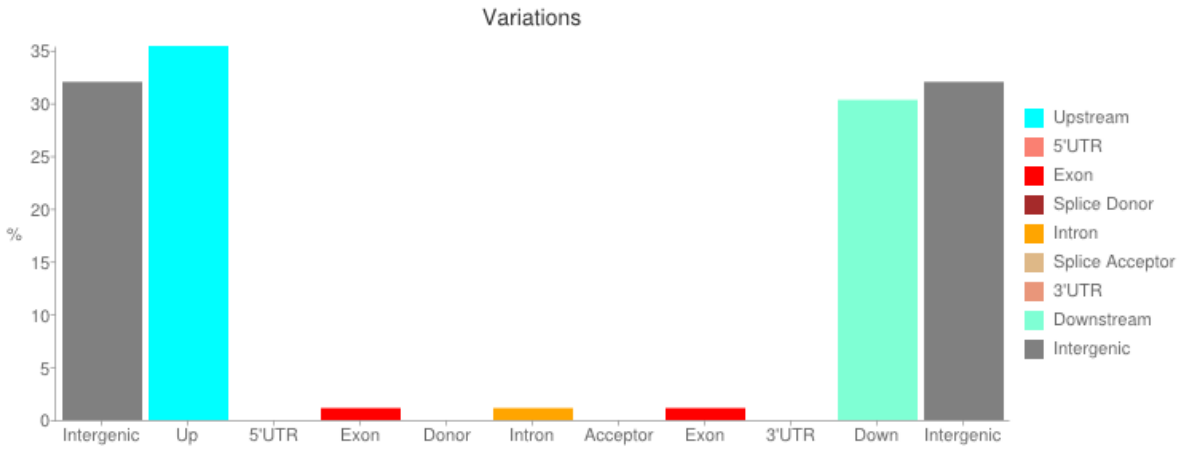


Figure 9: SNPs annotation to find the synonymous/non-synonymous effects of SNPs. It represents the distribution of SNPs in different regions of the aspen genome. It measures the percentage of SNPs (32%) observed in intergenic region, 35% SNPs were observed in upstream region, 30% SNPs were found in downstream region, whereas 2% of SNPs were observed in exonic region.

4. Discussion

The current project was conducted to assess two approaches that could be utilized for identifying a set of SNPs for later use to construct of a genetic map for *Populus tremula.*

Freeze dried leaves of aspen tree were used for DNA extraction. Restriction of aspen DNA was required to produce a reduced representation library of small fragments after expected restriction products. Initially, the expected quality and quantity of extracted DNA required for successful restriction analysis through restriction enzymes was not achieved. In the beginning aspen leaves were collected from open field conditions, which did not result in desired quality of DNA. The unexpected DNA quality, measured after repeated extractions may highlights the factor of growth stage of aspen leaves apart from other factors. It might also be speculated that use of different growth stages of aspen leaves perhaps result in variable qualities of the DNA. Restriction of aspen genomic DNA into small fragment size by restriction endonucleases was also difficult due to the leaf contamination from phenolic compounds. Repeated restriction analysis with 20 different restriction enzymes (Table 1) primarily did not result in desired restriction products of aspen DNA, which points towards failure in accessing the restriction sites of the aspen genome by those enzymes (Table 1). Later, DNA from fresh aspen and Arabidopsis leaves were extracted, which produced the expected bunch of small fragments ranging between 75-400 bp (Figure 4). Contaminations from phenolic compound along with fungal/yeast genomes diverted our future concentration from being continued with F1 population to switch towards SNP calling.

A huge pool of several thousand fragments on agarose gel was found around the range of 75-400 bp, which was also confirmed after *in silico* restriction analysis, where aspen DNA was restricted into a huge number of small fragments ranging 0-200bp (Figure 5-A, B and C). Presence of a heavy smear of fragments at low molecular weight might be due to the nature of *MseI* being a four base cutter enzyme (Table 1), as it further restricts the larger fragments produced by *PstI* in to smaller fragments when aspen and Arabidopsis genomic DNA were restricted with both of the enzymes together. Maximum numbers of small fragments are present in lane 1 and 3 (Figure 4), where DNA was fragmented by *MseI* alone and in combination with *PstI* respectively. However an opposite tendency of fragments was observed ranging 1000-10000 bp, where DNA was restricted by *PstI*. Restriction of aspen DNA into relatively larger fragments by *PstI* as compared to *MseI* is expected as it is a six base pair cutter, that is why a higher tendency of restricted fragments was observed in the

upper region of higher molecular weight (Figure 4 and 5-A, B and C). Genomic DNA is hard to restrict into small fragments as compared to plasmid DNA (Polisky and McCarty 1975). The reasons behind unexpected restriction might be the experimental errors, which highlights the quantity of DNA restricted by the enzymes, since with higher quantity (5.5 mg) of DNA, comparatively a desired pool of bands has been achieved (Tassell *et al.,* 2008). Restriction buffers were also properly vortexed to dissolve any precipitations. The samples were supplied sufficient time duration for restriction both for digestion at 37 ^{o}C and while gel running. Another option might be to use the polyacrylamide gel instead of agarose gel (Tassell *et al.,* 2008). Next-generation sequencing technologies have revolutionized the field of evolutionary biology, simplifying the genetic analysis relatively at larger scales. Further advancement in genetic analyses has been made by the advent of restriction site associated DNA (RAD) genotyping, a method that uses Illumina next-generation sequencing to discover and score tens to hundreds of thousands of single nucleotide polymorphism (SNP) markers in multiple individuals simultaneously (Etter *et al.,* 2011). The primary goal was to construct reduced representation library by RAD sequencing using genomic DNA of parental samples of aspen tree but on account of unexpected results by restriction analysis, the experimental strategy was swapped to use an alternative approach for the identification of homozygous SNPs *i.e.* RNA-Seq data for both male and female parents of aspen tree for homozygous SNP calling.

RNA-Seq data for both samples (229.1 and 349.2) from aspen with small fragment size (bp) of variable qualities were trimmed to remove bad quality sequence reads and only the good quality reads were selected for the *de novo* assembly to build a reference transcriptome. Since, sequencing of relatively smaller fragments might result in various sequencing errors, trimming of the sequences (Figure 6) through FastQ groomer and FastQC to avoid such errors help in selecting the good quality reads, which make the *de novo* assembly more reliable (Andrews 2012 and Sjödin *et al.,* 2009). Similarly, FastQ format stores biological sequence and its respective quality scores and it has been now become a de facto standard for storing the output of high throughput sequencing instruments such as Illumina Genome Analyzer (Sjödin *et al* 2009). The higher percentages (Table 3) of aligned reads for both samples (229.1 and 349.2) after trimming through FastQC further justify the application of FastQC and FastQ groomer.

To conduct the further analysis on sequencing data, sequence trimming after quality checking of reads helps in most reliable results (Goecks *et al* 2010). After trimming the length of sequence reads to 27-66 bp reads for both samples (Table 2), further analysis became easier

and produced reliable results. Relatively higher number of aligned reads, better per base sequence quality, improved per sequence quality scores, reduced number of N contents per base and zero (0) overrepresented sequences were the output after screening the reads through trimming, FastQ groomer, FastQC and Flagstat statistics. Screening of sequence reads for better quality also seem to be helpful in building the reference transcriptome sequence and later mapping these reads through BWA against the reference transcript and finally for SNP calling. Eliminating the proposed sequencing errors or bad quality reads also improved the per base sequence contents along with per base GC contents and per sequence GC contents, which further clarify our results (Table 2). Flagstat data clearly differentiates between aligned and unaligned reads, which also helps in avoiding the sequencing errors produced through low quality reads.

Rapid development in NGS technologies has created numerous challenges to handle the sequencing data, which has been solved by various computational tools, since "Trinity" provides a fast and successful algorithm used in *de novo* transcriptome assembly (Clarke *et al* 2013). Sequencing reads from sample 229.1 were selected for trinity assembly, where 95854 sequences were assembled together to build a reference transcriptome, and both the samples were mapped against this reference transcriptome. A reference transcript was needed to compare both the samples to see any difference between them at SNP level.

The assembled transcripts after Trinity were then aligned against *P.trichocarpa* genome with the help of an aligner "GMAP" since; GMAP provides an accurate gene structure with substantial polymorphism and sequencing errors (Wu and Watanabe 2005). Mapping the sequences against *P.trichocarpa* genome with GMAP was actually aimed to ensure the escape of mapping errors, which might be produced after BWA. After mapping with BWA the sequences obtained looked error prone, which were tried to remove after mapping with GMAP, since it provides higher quality alignments (Wu and Watanabe 2005).

Variation in genome sequences is the primary reason for keeping plant species diverse from others. The sequences of both samples were therefore targeted to locate the single nucleotide polymorphism. Genome Analysis Toolkit (GATK) was used for this purpose, which involves various steps starting from finding raw SNPs to end at extracting true SNPs positions in the genome. On account of facing sequencing errors coupled with allelic variant and copy number variants, the initial SNP data should be screened to separate real SNPs. GATK analysis tool kit provides with such initial steps of first realigning the sequence reads locally and then recalibrating them on the basis of base quality values (Appendix 3, Figure 13). The BAM

output files after each step are processed to accumulate the SNPs in the final step of GATK-Analysis, which stores raw SNPs in the VCF file. Fast Q quality control and flagstat basic statistics support the filtering and grooming of BAM files. GATK-Analysis screened out initially 8122 SNPs, which predicts presence of a single (1) SNP per 200 bp in the entire genome. Since sequence reads for both samples were compared for the presence of SNP, all the common transcripts showing SNPs in their sequences were extracted so that a comparative study could have been drawn. The process of SNPs filtering is further supported by sequencing errors, phenolic contaminations, genomic contaminations, non-allelic variants, poor SNP quality and low sequencing depth. Construction of a genetic map requires exact position of SNPs in the genome, which demanded us to identify where these SNPs exist in the entire *P.trichocarpa* genome sequence. The transcripts showing SNPs in their sequences from both of the samples were of different qualities (coverage) and sequencing depth, which require all the SNPs to be filtered based on SNP quality, sequencing depth (DP), allelic frequency (AF) and genotype information (GT) (Danecek *et al* 2011). The transcripts with relatively lower qualities were filtered out because these might be due to the allelic variants or sequencing errors instead of real SNPs. Identification of exact SNP position in reference transcript through IGV further validates the process of SNP calling, as it clearly represents both homozygous and heterozygous SNPs at different position in both samples and in the final SNP storing VCF file. Furthermore, the annotation of these SNPs also validates the type of effects these SNPs hold being synonymous or non-synonymous. Most of the filtered SNPs exist in intergenic region, which is highly variable, however most of the reads were aligned in the exonic region. Similarly, intergenic regions exist relatively away from the actual genes encoding proteins, which is very likely in case of RNA-Seq data (Figure 9).

Concludingly, since restriction analysis data of *Populus* genome was not reliable to analyse the F_1 and F_2 population, so the idea was shifted to SNP calling for homozygous SNP discovery in both of parental samples, which initially discovered 8122 SNPs in both parents. These 8122 SNPs were proposed to hold bad quality reads due to some sequencing errors, copy number variants and allelic variants. To obtain a high coverage SNPs data, the initial bunch of SNPs was filtered based on SNP quality and sequencing depth, where a threshold was fixed as 2.0-2.2 for SNP quality and 3.0-3.2 for sequencing depth. After applying filters over bad coverage SNPs, 118 SNP were discovered as good quality SNPs for chromosome 19 between both samples. However, validation of these putative SNPs by designing the allele specific and locus specific primers around the putative SNPs regions covering all 1-19

chromosomes would result in highly validated SNPs covering maximum coverage. Similarly, identification of synonymous and non-synonymous effects of all the 8122 SNPs would be a useful extension of this project. Analysis of F_1 and F_2 population for homozygous SNP discovery, enabling us to identify the inheritance pattern of parental genes would be another contribution to the aspen sequencing project. Furthermore, unrevealing unaligned transcripts by GMAP to *P. trichocarpa* would help us to extend our understanding about any possible *Populus*-microbial association since, there has a lot been reported about *Populus*-fungal genome reshuffling?

5. References

Adams M., D., Kelley J., M., Gocayne J., D., Dubnick M., Polymeropoulos M., H., Xiao H., Merril C., R., Wu A., Olde B. and Moreno R., F. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252(5013)**:**1651-1656.

Altshuler D., Pollara V. J., Cowles C. R., Van Etten W. J., Baldwin J., Linton L., Lander E.S. (2000). A SNP map of the human genome generated by reduced representation shotgun sequencing. Nature, 407(6803):513-516.

Andrews S. (2012). A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fasctqc.

Baginsky S., Hennig L., Zimmermann P. and Gruissem W. (2010). Gene Expression Analysis, proteomics, and Network Discovery. Plant PhysiologyD, Vol. 152, pp. 402–410,

Baird N. A., Etter P. D., Atwood T. S., Currey M. C., Shiver A. L., Lewis Z. A., Selker E. U., Cresko W. A. and Johnson E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. October, Volume 3, Issue 10 e3376.

Bao S., Jiang R., Kwan W., Wang B., Ma X. and Song Y. K. (2011). Evaluation of next-generation sequencing software in mapping and assembly. Journal of Human Genetics. 56, 406–414.

Barbazuk W. B., Bedell J. A. and Rabinowicz P. D. (2005). Reduced representation sequencing: a success in maize and a promise for other plant genomes. BioEssays 27:839–848.

Batzoglou S., Jaffe D. B. and Stanley K. (2002). "ARACHNE: a whole-genome shotgun assembler". Genome Res. 12 (1): 177–89.

Bozdag D., Barbacioru C. C. and Catalyurek, U. V. (2009). IEEE International Symposium on, Parallel & Distributed Processing, 1033–1042.

Bomar L., Maltz M., Colston S, and Graf J. (2011). Directed culturing of microorganisms using metatranscriptomics. mBio. 2:e00012–11.

Bouck A. and Vision T. (2007**).** The molecular ecologist's guide to expressed sequence tags. *Mol Ecol*, 16(5)**:**907-924.

Cervera M. T., Storme V., Ivens B., Gusma˜o J., Liu B. H., Hostyn V., Slycken J. V., Montagu M. V. and Boerjan W. (2001). Dense Genetic Linkage Maps of Three Populus Species (*Populus deltoides,P. nigra* and *P. trichocarpa*) Based on AFLP and Microsatellite Markers. Genetics 158: 787–809.

Cingolani P., Platts A., Wang le L., Coon M., Nguyen T., Wang L., Land S. J., Lu X., Ruden D. M. Fly (Austin). (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.", Apr-Jun; 6(2):80-92. PMID: 22728672.

Clarke K., Yang Y., Marsh R., Xie L. and Zhang K K. (2013). Comparative analysis of de novo transcriptome assembly. Life science. Vol.56 No.2: 156–162.

Cock P. J., Fields C. J., Goto N., Heuer M. L., Rice P. M. (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. Nucleic Acids Res. 2010, Apr;38(6):1767-71. doi: 10.1093/nar/gkp1137. Epub 2009.

Danecek P., Auton A., Abecasis G., Albers C. A., Banks E, et al. (2011) The variant call format and VCFtools. Bioinformatics 27: 2156–2158.

De Bruijn, N.G. (1946). A combinatorial problem. Koninklijke Nederlandse Akademie v. Wetenschappen 46, 758–764.

Depristo M., Banks E., Poplin R., Garimella K., Maguire J., Hartl C., Philippakis A., del Angel G., Rivas M. A., Hanna M., McKenna A., Fennell T., Kernytsy A., Sivachenko A., Cibulskis K., Gabriel S., Altschuler D. and Daly M. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 43:491-498.

Egan A. N., Schlueter J. and Spooner D. M. (2012). Applications of next-generation sequencing in plant biology. American Journal of Botany 99(2): 175–185.

Goecks, J., Nekrutenko, A., Taylor, J., Galaxy Team, T. (2010). "Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences". *Genome Biology* 11 (8): R86. doi:10.1186/gb-2010-11-8-r86.

Grabherr M. G., Haas B. J., Yassour M., Levin J. Z., Thompson D. A., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N., Palma F. D., Birren B. W., Nusbaum C., Toh K. L., Friedman N. and Regev A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology, volume 29 number 7 July.

Guttman M., Garber M., Levin J. Z., Donaghey J., Robinson J., Adiconis X., ,Fan L., Magdalena J Koziol M. J., Gnirke A., Nusbaum C., Rinn J. L., Lander E. S. and Regev A. (2010). Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrNAs. Nature biotechnology, Volume 28 number 5.

Gydle P. R. (2011). Next-generation gene catalogues and genomics tools focused on forestry research. IUFRO Tree Biotechnology Conference, "From genomes do integration and delivery" 26 June - 02 July, Arraial d´ajuda - BahiA - Brazil Conference Proceedings abstracts.

Haas, B. J. and Zody, M.C. (2010). Advancing RNA-Seq analysis. *Nat. Biotechnol.* 28, 421–423.

Horner, D. S., Pavesi, G., Castrignano, T., De Meo, P. D., Liuni, S., Sammeth, M. (2009). Bioinformatics approaches for genomics and post genomics applications of next generation sequencing. Brief. Bioinform. 11, 181–197.

Hyten D., Cannon S. B, Song Q., Weeks N., Fickus E. W., Shoemaker R. C., Specht J. E., Farmer A. D., May G. D. and Cregan P. B. (2010). High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. BMC Genomics, 11:38.

Imbert, E., Lefèvre, F. (2003). Dispersal and gene flow of *Populus nigra* (Salicaceae) along a dynamic river-system. Jour. Ecol. 91: 447-456.

Kim E. C., Lee H. S. and Choi D. W. (2012). Sequence variability and expression pattern of the dehydrin gene family in *Populus alba* × *P. tremula* var. glandulosa. POJ 5(2):122-127.

Li H. and Durbin R. (2010) Fast and accurate long-read alignment with Burrows-

Li H., Ruan J. and Durbin R. (2012). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18:1851-8.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics. 15;25(14):1754-60.

Liu, Z., and Furnier G. R. (1993). Comparison of allozyme, RFLP, and RAPD markers for revealing genetic variation within and between trembling aspen and bigtooth aspen. Theor. Appl. Genet. 87: 97–105

Luca F., Hudson R. R., Witonsky D. B. and Rienzo A. D. (2012). A reduced representation approach to population genetic analyses and applications to human evolution. Genome Research 21:1087–1098.

Mazur, B. J., and S. V. Tingey. (1995). Geneticmapping and introgres sion of genes of agronomic importance. Curr. Opin. Biotechnol. 6: 175–182.

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altschuler D., Gabriel S., Daly M. and DePristo M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297-303.

Metzker M L. (2010). Sequencing technologies —the next generation. Nature Reviews Genetics volume 11, January 2010.

Mu¨ller-Starck G. (1992). Genetic control and inheritance of isoen zymes in poplars of the Tacamahaca section and hybrids. Silvae Genet. 41: 87–95. Nat. Biotechnol. 27, 455–457.

Paolucci I., Gaudet M., Jorge V., Beritognolo I., Terzoli S., Kuzminsky E., Muleo R., Mugnozza G. S. and Sabatti M. (2010). Genetic linkage maps of Populus alba L. and comparative mapping analysis of sex determination across Populus species. Tree Genetics & Genomes, DOI 10.1007/s11295-010-0297-7.

Paszkiewicz K. and Studholme D. J. (2010). De novo assembly of short sequence reads. Brief Bioinform.11(5):457-72. doi: 10.1093/bib/bbq020.

Polisky B. and McCarty B. (1975). Location of histones on simian virus 40 DNA. Proc. Nat. Acad. Sci. USA, Vol. 72, No. 8, pp. 2895-2899.

Pop M. (2009). Genome assembly reborn: recent computational challenges, Briefings in bioinformatics. VOL 10. NO 4. 354 -366.

Qiu Q., Ma T., Hu Q., Liu B., Wu Y., Zhou H., Wang Q., Wang J. and Liu J. (2010). Genome-scale transcriptome analysis of the desert poplar, Populus euphratica. Tree Physiology Online at http://www.treephys.oxfordjournals.org.

Rinaldi C, Kohler A, Frey P, Duchaussoy F, Ningre N, Couloux A, Wincker P, Le Thiec D, Fluch S, Martin F, Duplessis S. (2007). Plant Physiol;144(1):347-66.

Sanger F., Nicklen S., Coulson A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 1977; 74:5463–67.

Shendure J and J Hanlee. (2008). Next-generation DNA sequencing. Nature biotechnology, volume 26 number 10, 1135-1145.

Simpson J. T., Wong K., Jackman S. D., Schein J. E., Jones S. J. M. and Birol I. (2009). ABySS: A parallel assembler for short read sequence data. Genome Research. 19:1117–1123.

Sjödin A, Street NR, Sandberg G, Gustafsson P and Jansson S (2009) PopGenIE: The Populus Genome Integrative Explorer. A new tool for exploring the Populus genome. New Phytologist 2009.

Stevens, M.T., Turner, M.G., Tuskan, G.A., Romme, W.H., Gunter, L., Waller, D.M. (1999). Genetic variation in postfire aspen seedling in Yellowstone National Park. Mol. Ecol. 8: 1769–1780.

Tassell C. P. V.., Smith T. P. L., Matukumalli L. K., Taylor J. F., Schnabel R. D., Lawley C. T., Haudenschild C. D., Moore S. S., Warren W. C. and Sonstegard T. S. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature methods VOL.5 NO.3.

Trapnell C. and Salzberg, S. L. (2009). How to map billions of short reads onto genomes. *Nat Biotechnol*; 27(5): 455–457.

Trapnell, C., Roberts A., Goff L., Pertea G., Kim D., Kelley DR, Pimentel H., Salzberg SL,Rinn JL, and Pachter L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols, 562-578.

Tuskan, G. A., Difazio, S. and Jansson, S. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science, 313, 1596–1604.

Wu T. D. and Watanabe C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 1;21(9):1859-75.

Yeh, F.C., Chong, D.K.X., Yang, R.C. (1995). RAPD variation within and among natural populations of trembling aspen (*Populus tremuloides* Michx.) from Alberta. J. Hered. 86: 454–460.

Yin, T. M., DiFazio, S. P., Gunter, L. E., Riemenschneider, D. and Tuskan, G.A. (2004). Large-scale heterospecific segregation distortion in Populus revealed by a dense genetic map. Theor. Appl. Genet. 109, 451–463.

Young . L., H. O., Mullikin J. C., E. and Margulies E. H. (2010). A new strategy for genome assembly using short sequence reads and reduced representation libraries. Genome Res. 20: 249-256.

Zerbino, D.R. and Birney, e. Velvet. (2008). Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18, 821–829.

6. Supplementary data

APPENDIX 1

Table 4: Qubit Protocol

1. Set up two assay tubes for the standards (three for the protein assay) and one tube for each user sample.
2. Prepare the Qubit **working solution** by diluting the Qubit reagent 1:200 in Qubit buffer. Prepare 200 µl of **working solution** for each standard and sample.
3. Prepare the Assay tubes according to the table below.

	Standard assay tubes	**User sample** assay tubes
Volume of **working solution** (from step 2) to add	190 µl	180-199 µl
Volume of **standard** (from kit) to add	10 µl	-
Volume of **user sample** to add	-	1-20 µl
Total volume in each assay tube	200 µl	200 µl

4. Vortex all tubes for 2-3 seconds.
5. Incubate the tubes for 2 minutes at room temperature (15 minutes for the Qubit protein assay)
6. Insert the tubes in the Qubit 2.0 Fluorometer and take readings. For detailed instructions, refer to the Qubit 2.0 Fluorometer manual.
7. Optional: using the dilution calculator feature of the Qubit 2.0 Fluorometer, determine the stock concentration of your original sample.

APPENDIX 2

Supplementary Gel images of restriction analysis

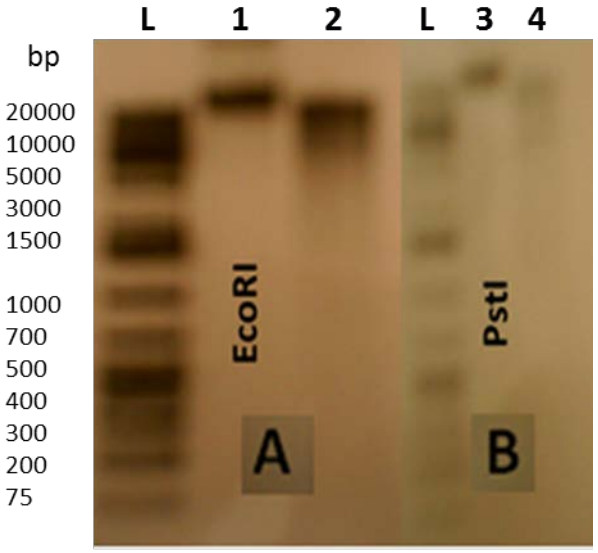


Figure 10: Restriction analyses of aspen DNA by different individual restriction enzymes. It shows restriction analysis of aspen genome by restriction enzyme *EcoRI* **(A)** and Arabidopsis genome by *PstI* **(B)**. **L**= ladder, **1 and 3**= Positive control (undigested DNA) from male parent and **2 and 4**= Experimental sample (digested DNA) from male parent.

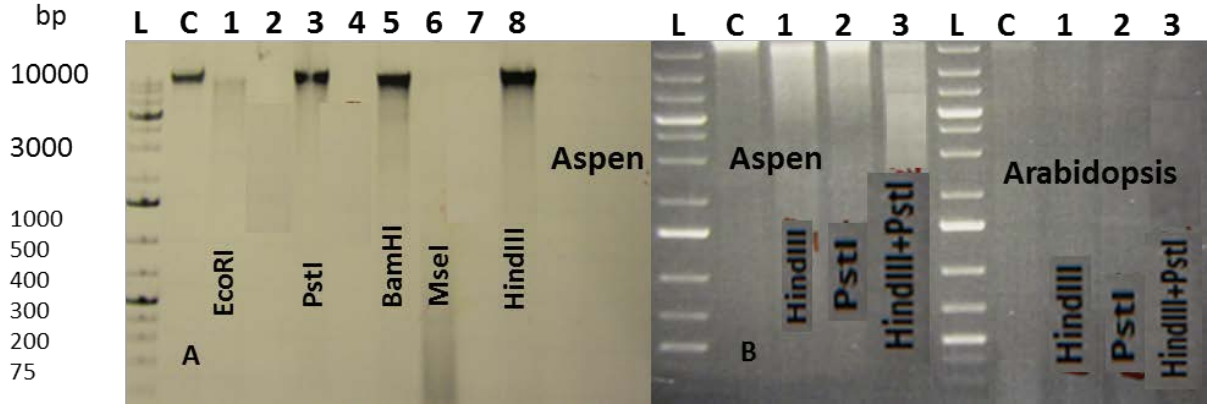


Figure 11: Restriction analysis of aspen DNA by combination of different restriction enzymes. Restriction analyses of aspen DNA by combinations of different restriction enzymes. **A** represents the banding pattern by *EcoRI, PstI, BamHI, MseI* and *HindIII* respectively. **B** represents the combination of two restriction enzymes (*PstI* and *HindIII*) in aspen genome ($\mathbf{L^P}$) and Arabidopsis ($\mathbf{L^A}$) genome. **L**= ladder, **C**= Control sample, $\mathbf{L^P}$= Aspen DNA and $\mathbf{L^A}$= Arabidopsis DNA.

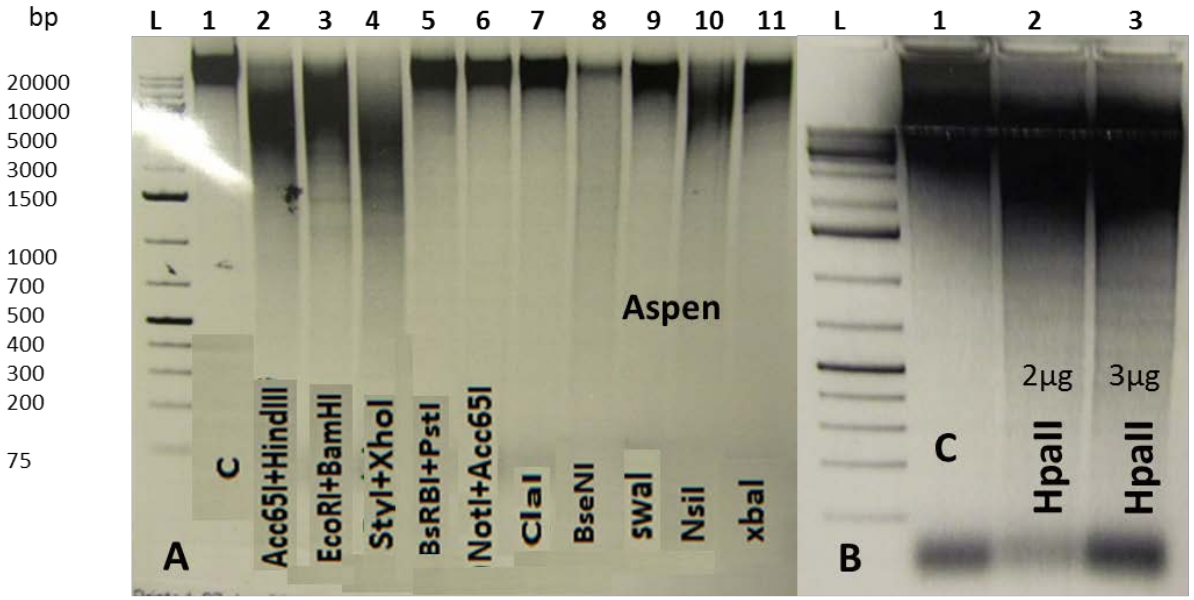
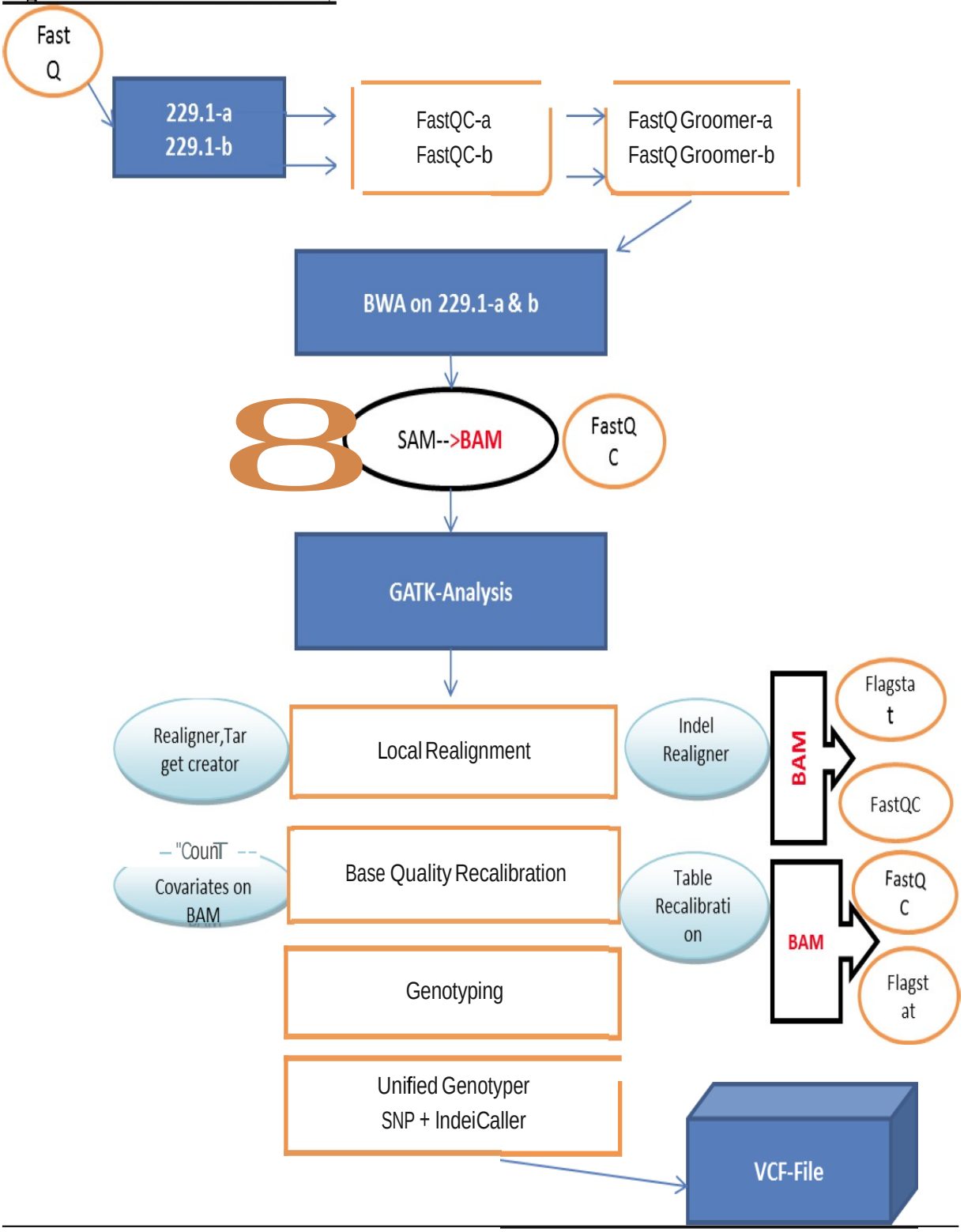


Figure 12: Restriction analyses of aspen DNA by combinations of different restriction enzymes. **A** shows restriction analysis by single enzyme (lane # 6,7,8,9 and 10) and by combination of two enzymes ((lane # 1,2,3,4 and 5), **B** indicates restriction analysis of *HpaII*. **L**= ladder, **C**= Positive control sample.

APPENDIX3

Figure 13: Workflow on Galaxy



APPENDIX 4

Table 5: A subset of high quality 118 SNPs

Sr.No.	Transcript IDs	Start position	End position	Allele	Genomic position	SNP quality	Depth	log10 (qual)	log10 (depth)
1	comp32610_c0_seq3	823	824	A/T	2284166	1579,87	112	3,1986214	2,04921802
2	comp26364_c0_seq1	474	475	A/T	14610163	1573,19	118	3,1967812	2,07188201
3	comp30658_c1_seq9	46	47	C/T	13311283	1563,58	102	3,1941201	2,00860017
4	comp26364_c0_seq1	200	201	G/T	14609889	1552,32	121	3,1909813	2,08278537
5	comp32184_c2_seq1	1522	1523	G/C	10797327	1540,77	120	3,1877378	2,07918125
6	comp30150_c1_seq1	921	922	C/T	8752245	1520,81	115	3,182075	2,06069784
7	comp32931_c0_seq1	1071	1072	G/A	15378250	1520,24	135	3,1819122	2,13033377
8	comp25325_c3_seq1	1753	1754	A/G	10492753	1506,14	117	3,1778653	2,06818586
9	comp30597_c0_seq1	256	257	G/A	11637549	1505,2	102	3,1775942	2,00860017
10	comp29409_c2_seq2	828	829	G/A	12156336	1498,26	103	3,1755872	2,01283722
11	comp30647_c0_seq1	212	213	T/A	10543527	1489,38	123	3,1730055	2,08990511
12	comp33175_c2_seq5	85	86	C/T	344902	1488,07	103	3,1726234	2,01283722
13	comp32121_c0_seq1	3569	3570	T/G	15315080	1483,76	154	3,1713637	2,18752072
14	comp30150_c1_seq1	735	736	A/G	8752059	1481,58	125	3,1707251	2,09691001
15	comp30647_c0_seq1	208	209	A/C	10543523	1481,55	122	3,1707163	2,08635983
16	comp30527_c0_seq1	760	761	T/C	1545347	1473,09	152	3,1682293	2,18184359
17	comp31561_c0_seq28	235	236	T/C	14766457	1465,4	152	3,1659562	2,18184359
18	comp32184_c2_seq1	1533	1534	A/G	10797338	1463,47	111	3,1653838	2,04532298
19	comp30647_c0_seq1	172	173	A/T	10543487	1462,54	121	3,1651078	2,08278537
20	comp31002_c1_seq2	371	372	G/A	14907989	1458,41	126	3,1638796	2,10037055
21	comp31265_c0_seq5	454	455	C/A	11571730	1455,9	156	3,1631315	2,1931246
22	comp33400_c0_seq5	895	896	T/C	4568682	1454,73	107	3,1627824	2,02938378
23	comp20667_c0_seq1	390	391	C/G	15700891	1445,61	156	3,1600511	2,1931246
24	comp32348_c2_seq1	1064	1065	A/G	11492565	1438,97	139	3,1580517	2,1430148
25	comp33307_c6_seq2	374	375	G/A	2185912	1437,9	107	3,1577287	2,02938378
26	comp20667_c0_seq1	381	382	G/C	15700882	1433,75	158	3,1564734	2,19865709
27	comp26364_c0_seq1	372	373	T/A	14610061	1420,57	149	3,1524626	2,17318627
28	comp26372_c0_seq2	214	215	T/G	13780762	1405,62	145	3,1478679	2,161368
29	comp33258_c0_seq7	280	281	A/T	2152384	1405,22	154	3,1477443	2,18752072
30	comp22272_c0_seq1	476	477	T/G	10526674	1386,85	157	3,1420295	2,19589965
31	comp26416_c0_seq1	1736	1737	G/A	247931	1378,31	151	3,1393469	2,17897695
32	comp32044_c1_seq2	185	186	A/T	3901599	1374,9	111	3,1382711	2,04532298
33	comp32931_c0_seq1	1083	1084	C/T	15378262	1371,12	144	3,1370755	2,15836249
34	comp32931_c0_seq1	618	619	T/A	15377797	1362,81	147	3,1344353	2,16731733
35	comp24841_c0_seq1	173	174	A/G	9386524	1352,58	153	3,131163	2,18469143
36	comp29921_c0_seq1	732	733	G/T	10199123	1350,66	133	3,130546	2,12385164
37	comp17541_c1_seq1	397	398	T/A	10649119	1349,44	120	3,1301536	2,07918125
38	comp20667_c0_seq1	383	384	A/C	15700884	1343,81	150	3,1283379	2,17609126
39	comp33307_c6_seq2	275	276	C/T	2185813	1341,97	104	3,1277428	2,01703334
40	comp24841_c0_seq1	14	15	G/A	9386365	1338,79	115	3,1267125	2,06069784
41	comp32931_c0_seq1	906	907	C/A	15378085	1332,54	147	3,1246803	2,16731733
42	comp32931_c0_seq1	1079	1080	A/G	15378258	1331,79	143	3,1244357	2,15533604
43	comp23466_c0_seq1	54	55	T/G	3945861	1328,49	102	3,1233583	2,00860017
44	comp26364_c0_seq1	359	360	C/A	14610048	1322,69	146	3,1214581	2,16435286
45	comp33258_c0_seq6	712	713	A/G	2146201	1297,42	118	3,1130806	2,07188201
46	comp29281_c2_seq1	322	323	A/G	8542046	1289,73	156	3,1104988	2,1931246
47	comp31192_c1_seq3	1435	1436	T/G	15163251	1287,72	145	3,1098214	2,161368
48	comp23466_c0_seq1	49	50	G/A	3945856	1285,88	100	3,1092004	2
49	comp26364_c0_seq1	466	467	A/T	14610155	1274,75	105	3,105425	2,0211893
50	comp31325_c0_seq2	776	777	T/G	7783108	1273,67	132	3,1050569	2,12057393
51	comp32373_c1_seq9	210	211	T/C	12178250	1241,11	147	3,0938103	2,16731733
52	comp26364_c0_seq1	346	347	G/A	14610035	1216,69	139	3,0851799	2,1430148
53	comp20245_c0_seq1	343	344	A/G	4852609	1214,31	140	3,0843296	2,14612804
54	comp28577_c1_seq2	629	630	T/A	150934	1213,51	140	3,0840434	2,14612804
55	comp31561_c0_seq28	264	265	A/T	14766486	1210,81	141	3,083076	2,14921911
56	comp26671_c0_seq1	401	402	T/A	3989674	1204	149	3,0806265	2,17318627
57	comp30659_c0_seq3	891	892	G/T	220911	1200,37	123	3,0793151	2,08990511
58	comp29356_c0_seq1	1288	1289	C/A	10672997	1197,53	152	3,0782864	2,18184359
59	comp26308_c0_seq1	174	175	G/A	4033935	1195,97	117	3,0777203	2,06818586
60	comp32219_c0_seq2	393	394	C/G	11564417	1195,81	103	3,0776622	2,01283722
61	comp26659_c0_seq1	242	243	G/A	15446323	1190,97	134	3,0759008	2,1271048
62	comp31325_c0_seq2	781	782	T/G	7783113	1190,61	134	3,0757695	2,1271048
63	comp32219_c0_seq2	279	280	C/T	11564303	1182,39	118	3,0727607	2,07188201

64	comp29032_c0_seq2	274	275	A/G	11135259	1180,5	119	3,072066	2,07554696
65	comp32121_c0_seq1	2838	2839	C/A	15314349	1174,38	130	3,0698086	2,11394335
66	comp26316_c0_seq1	681	682	C/G	11630212	1170,07	158	3,0682118	2,19865709
67	comp28577_c1_seq2	735	736	T/C	151040	1162,31	108	3,065322	2,03342376
68	comp23466_c0_seq1	345	346	C/T	3946152	1160,34	123	3,0645853	2,08990511
69	comp29150_c0_seq1	161	162	A/T	11155065	1159,68	120	3,0643382	2,07918125
70	comp32121_c0_seq1	3910	3911	T/A	15315421	1157,8	134	3,0636335	2,1271048
71	comp29150_c0_seq1	244	245	A/G	11155148	1156,24	142	3,063048	2,15228834
72	comp28577_c1_seq2	1919	1920	A/G	152224	1154,82	128	3,0625143	2,10720997
73	comp24841_c0_seq1	18	19	C/T	9386369	1150,32	143	3,0608187	2,15533604
74	comp33413_c1_seq4	1582	1583	G/A	2924863	1149,32	107	3,060441	2,02938378
75	comp29150_c0_seq1	243	244	A/G	11155147	1148,35	141	3,0600743	2,14921911
76	comp32931_c0_seq1	1023	1024	A/G	15378202	1145,59	136	3,0590292	2,13353891
77	comp32094_c0_seq1	538	539	G/C	10715287	1145,41	121	3,058961	2,08278537
78	comp27363_c0_seq2	524	525	G/T	11500252	1139,51	100	3,0567181	2
79	comp30153_c2_seq1	1332	1333	C/G	8870133	1129,93	133	3,0530515	2,12385164
80	comp32931_c0_seq1	1027	1028	A/T	15378206	1123,13	141	3,05043	2,14921911
81	comp31899_c3_seq14	104	105	T/G	7263974	1114,68	122	3,0471502	2,08635983
82	comp82519_c0_seq1	292	293	G/T	12380453	1111,49	115	3,0459056	2,06069784
83	comp30659_c0_seq3	691	692	A/T	220711	1108,44	134	3,0447122	2,1271048
84	comp32632_c0_seq8	304	305	C/T	15671067	1106,51	108	3,0439553	2,03342376
85	comp32931_c0_seq1	858	859	A/G	15378037	1106,41	117	3,0439161	2,06818586
86	comp31265_c0_seq5	340	341	A/G	11571616	1105,98	137	3,0437473	2,13672057
87	comp29150_c0_seq1	160	161	A/T	11155064	1105,81	119	3,0436805	2,07554696
88	comp26372_c0_seq2	224	225	G/T	13780772	1101,29	118	3,0419017	2,07188201
89	comp28577_c1_seq2	329	330	C/A	150634	1099,95	118	3,0413729	2,07188201
90	comp31389_c0_seq2	1913	1914	T/C	15197857	1096,62	142	3,0400562	2,15228834
91	comp23466_c0_seq1	318	319	T/C	3946125	1094,96	135	3,0393983	2,13033377
92	comp28577_c1_seq2	149	150	A/T	150454	1090,43	105	3,0375978	2,0211893
93	comp33332_c0_seq2	155	156	A/G	14675440	1087,53	113	3,0364412	2,05307844
94	comp33441_c3_seq1	2799	2800	T/C	15760214	1080,55	155	3,0336449	2,1903317
95	comp32348_c2_seq1	587	588	A/G	11492088	1077,14	148	3,0322722	2,17026172
96	comp33392_c1_seq4	359	360	T/C	7924579	1076,36	155	3,0319576	2,1903317
97	comp26548_c0_seq1	551	552	G/T	8479558	1073,14	108	3,0306564	2,03342376
98	comp29032_c0_seq5	46	47	C/G	11131278	1072,49	148	3,0303933	2,17026172
99	comp33392_c1_seq4	257	258	T/G	7924477	1070,14	155	3,0294406	2,1903317
100	comp26397_c0_seq1	70	71	T/G	12898384	1066,35	118	3,0278998	2,07188201
101	comp32044_c1_seq2	122	123	T/C	3901536	1064,91	100	3,0273129	2
102	comp31689_c6_seq1	321	322	C/T	4973673	1064,4	139	3,0271049	2,1430148
103	comp30150_c1_seq1	768	769	A/T	8752092	1057,78	136	3,0243954	2,13353891
104	comp31639_c5_seq2	301	302	G/A	13037934	1056,37	107	3,0238161	2,02938378
105	comp29150_c0_seq1	364	365	T/C	11155268	1053,26	115	3,0225356	2,06069784
106	comp30597_c0_seq1	875	876	G/A	11638168	1040,13	105	3,0170876	2,0211893
107	comp33392_c1_seq4	395	396	A/C	7924615	1038,8	136	3,0165319	2,13353891
108	comp31689_c6_seq1	375	376	C/G	4973727	1033,99	122	3,0145163	2,08635983
109	comp33392_c1_seq4	431	432	T/C	7924651	1033,92	133	3,0144869	2,12385164
110	comp33258_c0_seq6	679	680	G/T	2146168	1030,18	115	3,0129131	2,06069784
111	comp30527_c0_seq1	129	130	A/G	1544716	1027,79	104	3,0119044	2,01703334
112	comp32121_c0_seq1	3116	3117	G/A	15314627	1025,89	123	3,0111008	2,08990511
113	comp33392_c1_seq4	269	270	T/A	7924489	1024,44	155	3,0104865	2,1903317
114	comp32106_c1_seq5	1497	1498	C/T	10625960	1016,31	142	3,0070262	2,15228834
115	comp32931_c0_seq1	1005	1006	T/C	15378184	1010,37	147	3,0044804	2,16731733
116	comp33441_c3_seq1	2774	2775	T/C	15760189	1006,55	146	3,0028354	2,16435286
117	comp28577_c1_seq2	568	569	T/C	150873	1006,08	125	3,0026325	2,09691001
118	comp33413_c1_seq4	1612	1613	T/A	2924893	1004,18	113	3,0018116	2,05307844

7. List of abbreviations

NGS	**Next Generation Sequencing**
GMAP	**Genomic Mapping Alignment Program**
BWA	**Burrows Wheeler Alignment**
GATK	**Genomic Analysis Toolkit**
RAD	**Restriction Associated DNA**
TAIR	**The Arabidopsis Information Resource**
DNA	**Deoxyribonucleic Acid**
RNA	**Ribonucleic Acid**
SNP	**Single Nucleotide Polymorphism**
SAM	**Sequence Alignment Map**
BAM	**Binary Alignment Map**
VCF	**Variant Call Format**
BED	**Browser Extensible Data**
GFF	**General Feature Format**
IGV	**Integrative Genome Viewer**
ID	**Identity Number**
DP	**Sequencing Depth**
GT	**Genotypic Information**
AF	**Allelic Frequency**
QUAL	**Quality value**
RRL	**Reduced Representation Library**
EST	**Expressed Sequence Tags**